AWARD NUMBER:
W81XWH-13-1-0061

TITLE:
Novel Visualization of Large Health Related Data Sets

PRINCIPAL INVESTIGATOR:
William Ed Hammond, PhD

CONTRACTING ORGANIZATION:
Duke University
Durham, NC 27706

REPORT DATE:
March 2015

TYPE OF REPORT:
Annual

PREPARED FOR:
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:
Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
March 2015	Annual Technical Report	25 Feb 2014 - 24 Feb 2015

4. TITLE AND SUBTITLE

Novel Visualization of Large Health Related Data Sets

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-13-1-0061

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Hammond, William, E; West, Vivian; Borland, David; Akushevich, Igor; Heinz, Eugenia, McPeck
email: william.hammond@dm.duke.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University
2200 W. Main St, Ste 710
Durham, NC 27705-4677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

USA Med Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT Using retrospective data queries to understand what information clinicians seek from health care data, we have identified data elements and are looking at combinations of data elements used in queries. We are developing various visualization techniques that can be used to present the informational content in large databases, expecting that visualization of this data will present or "discover" information without specific hypotheses. Groups of related data elements are incorporated into visualizations that allow a quick comparison of data from a large population, with the ability to view trends over time within a chosen category. We are exploring the ability to compress petabytes of health care data representing many data elements into various groups of related data presented visually with an interface that allows the user to interactively explore the data elements to understand big data from the perspective of an entire population, different disease groups, ages, and other variables. There is the potential to detect causal relationships between various sets of data, which when applied to military EHR data may lead to improved health care and resiliency in military personnel, assist the DoD in strategic decisions related to personnel, and save millions of dollars in health care costs.

specific aims, results of findings and their significance, and plans for the coming year

15. SUBJECT TERMS
Visualization, health care data, big data

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUM OF PAGES	
a. REPORT	b. ABSTRACT	c. THIS PAGE			19a. NAME OF RESPONSIBLE PERSON: USAMRMC
U	U	U	UU	63	19b. TELEPHONE NUMBER (include area code)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

1. INTRODUCTION

With the growth of Electronic Health Record (EHR) data and other related health care databases, there is a need to understand what information and knowledge the data represent. Visualization offers an opportunity to explore and understand large amounts of data in unique and novel ways, permitting one to view data without the bias of an a priori decision of what is important. We hypothesize that data visualization is more effective than traditional methods of data exploration, and that the type of visualization is highly dependent on the types of data and nature of the queries and what someone is trying to learn from the data. With TATRC support, we have used retrospective data queries at Duke University to understand what information clinicians seek from health care data, identify what data elements and mixtures of data classes (laboratory data, medications, diagnoses, therapies, demographic data, problems, physical examination data, or imaging data) are used in queries and what methods are used to analyze query results. We have applied data visualization methods with specific data elements from multiple classes to test our hypothesis and have reported some of our results at the annual American Medical Information Association (AMIA) meeting (see Appendix). We are now beginning to test visualization of mixed data classes.

2. BODY

Funding for this research began on February 24, 2013. We were unable to begin working with actual clinical data, however, until the end of September 2013 due to a delay in approval from the Human Research Protection Office (HRPO). Our timeline for completing our project milestones and associated tasks by August 24, 2014, the date we were to complete this research, was therefore delayed. We requested and received a no-cost extension to August 24, 2015 to complete the project as planned using a slightly revised timeline. The milestones in this report are based on our revised timelines.

Research accomplishments associated with the tasks within each of our nine milestones are detailed in 2.1 through 2.9. Section 2.10 describes future work to complete tasks and problems encountered.

2.1. Obtain access to queries of Duke's Clinical Data Warehouse. STATUS of Milestone: Completed.

At Duke we use an on-line query system called DEDUCE (Duke Enterprise Data Unified Content Explorer) to access data in the Decision Support Repository, which consists of hundreds of tables, some with hundreds of millions of rows of data, collected from over 3.2 million patients at Duke. DEDUCE was operationalized in 2008 and has been upgraded numerous times since its inception.[1]

DEDUCE is recognized within the Duke medical community for its value in abstracting information from the Data Repository, with users able to query approximately 10,000 data elements and refine the query to facilitate exploration of aggregate clinical data in support of operations, quality, and research. Output from the queries is in the form of common-separated values (CSV) files, ASCII files, Excel files, or simple graphs. Every query is saved on a Duke server. We requested and have been working with the retrospective query data from 01 Jan 2011 through 31 Jul 2013.

2.2. Develop classification for queries: STATUS: Completed.

- **Identify early use of data queries;**
- **Based on the reason for the queries, group them accordingly;**
- **Obtain access to AHLTA de-identified data, and using work from the Duke queries and classes, compare for similarities and differences and revise classes as needed.**

Last year we conducted a survey of DEDUCE users to identify what data users we seeking. Results of this survey were reported in our 2014 Annual Report.

After an exhaustive review of the retrospective DEDUCE data queries and their structure, we concluded that the queries are best organized in two broad groups: those requesting information for research (with numerous variations of the type of research and purpose, e.g. cohort identification or counts of patients to determine feasibility of participation in a clinical trial) and those seeking information for quality improvement. Randomly investigating the data elements proved to be overwhelming, as the amount of data in the queries is immense, making a manual review beyond that we did unrealistic.

The visualization of DEDUCE we completed as we began this research is perhaps the most useful way to understand the data elements used in the queries. Using a force-directed network visualization, we used counts of the various data elements queried using DEDUCE and the users to interactively explore the departments conducting queries, the data elements that were queried, and the data elements that were used most frequently in addition to their correlations. We found that data queries for research were the most frequent types of queries, with ICD9 codes and dates of each encounter the most frequently requested data elements.

We also validated the importance of interactive visualization: visualizing the data provides more information than a manual review of large amounts of data. The complexity of the data does require the use of additive measures of points of interest. To see various correlations and highlight the relationships among the data elements the visualization must provide a way to interactively highlight certain types of data to. A static view prevents a clear understanding of the data being visualized.

There is a learning curve for users of DEDUCE, and the query tool constrains the types of data that can be queried. It is obvious that more seasoned users of the system query more data than newer users, which may be a result of becoming more familiar with the tool itself. The retrospective queries we looked at were through July 2013. Since this time, revisions and improvements to DEDUCE have continued and it is possible that retrospective data from July 2013 to the present might provide different results.

We had planned on using data from AHLTA to compare the Duke queries and classes for similarities and differences. We were informed last year that AHLTA data would not be available for our use, however. We obtained a synthetic dataset of 1 million patients developed for the DoD, but found the data unsuitable. For example, everyone is listed with their home address around the Bethesda area, and data values for HgA1c are the same. We talked to the developers of the data and found that data are constructed based on input from medical experts who provide the developers with likely symptoms, diagnoses, and therapies. The developers can

customize the system to generate customized data for us, but it is a rules-based generation of data. Because our research is based on the premise that visualization a priori of large data will lead to discoveries not previously known, the synthetic data does not lend itself to our research. Therefore, we continue to use data from the Duke Data warehouse and run queries on this data to visualize and discover new knowledge. Our original plan was to use AHLTA data to look at people with a diagnosis of PTSD; we are now working with PTSD patient data from Duke

2.3. Identify data elements used in queries: Identify which queries should be most meaningful to include in our analysis and the individual researchers associated with those queries; and Group data elements into classes (e.g. laboratory data, demographics, medications). STATUS: Completed.

There are over 10,000 data elements that can be queried using DEDUCE and almost 600 users. We had originally intended to group queries by the types of researchers running queries on the data warehouse and to evaluate associated data elements to assist us as we asked researchers to evaluate the usefulness of various types of visualization. Many of the queries in our retrospective data were not actual queries, but tests by the developers of DEDUCE as they revised and upgraded the tool. We also found repeated similar queries by the same person. Pediatricians asked for ages in months, and most researchers wanted to know the ICD9 codes and dates of encounters, or visits. Other than this, there are 500 data elements—those available to researchers prior to adding geospatial data, which happened after July 2013, the end date of retrospective queries we have been using. A random selection of retrospective queries found most to be used. Data from several queries has been used to develop both standard and more advanced visualizations, and we are in the process of formalizing these to be used in interviews, described in 2.4 below.

As we conducted our own queries, we downloaded results from a query more than once as we had questions about different data elements, which is most likely the reason for the repeated similar queries by the same person. Because all DEDUCE queries are saved for an audit trail, there are many variable related to the queries and data elements. Without knowing the reason for each query, we cannot tell why any of the data elements were selected. We will include this question in our interview schedule for interviews we will conduct next quarter.

2.4. Explore alternative visualization methods of the data. Clinicians will use the Follow-up Questionnaire to compare alternate visualization of the data to the original presentation of the query data: STATUS: work is ongoing.

- **Explore visualization methods previously applied to health care;**
- **Conduct semi-structured interviews with clinicians to determine how the user intended to use the data from the query, the relevance of the query, and the clinicians' satisfaction and use of the information derived by the query;**
- **Develop visualizations of retrospective query data using standard visualization techniques and novel visualization techniques to share with clinicians; develop visualizations of retrospective query data using standard visualization techniques and novel visualization techniques to share with clinicians;**

- **Develop a Follow-up Questionnaire using a 5 point Likert scale to be used in researcher evaluation of different visualization methods; and**
- **Combine interview data with Questionnaire results to evaluate clinical relevance of the visualization methods.**

A systematic literature review of visualization methods previously used in health care was conducted in the spring 2013. Results from this review were published online first in JAMIA.[2] (See Appendix for copy) The journal publication is expected by the end of March 2015.

As we waited for HRPO approval to use data, we conducted an online survey instead of conducting interviews as planned and reported results in last year's Annual Report. We also asked the users if they would be interested in working with us further on this project. Of the 61 responders to the survey, there were 34 people who provided us with contact information and indicated a willingness to be contacted individually, and are in the process of setting up interviews with these people.

Several visualizations have been developed with select retrospective query data, including bar charts, heat maps, and tree maps. We are in the process of developing radial coordinates, parallel sets, and path map visualizations using the same data. We expect to have a portfolio of various visualizations for evaluation during semi-structured interviews with respondents. Using REDCap, we are finalizing a questionnaire that will be used to by the respondents to evaluate each visualization method using a 5-point Likert scale.

2.5. Modify or revise classification and data elements of queries based on analysis of the relevance of the visualization methods. STATUS: Work has not yet begun, pending completion of Milestone 4.

2.6. Create a matrix of best visualization techniques. STATUS: A matrix of best visualization techniques will be completed when we have clinician input regarding the various techniques used.

2.7. Explore ways to mix different types of data in visualization. Develop parallel-coordinates visualization of data. STATUS: Work is ongoing.

We have continued exploring ways of mixing different types of data, including improving the integration of numerical and categorical data in our Radial Coordinates visualization, integrating temporal and categorical information in our visualization of diabetes-related data, and experimenting with the visualization of textual information from the literature related to visualization in health care.

Radial Coordinates

To enhance the ability to perceive relationships with categorical axes in our radial coordinates visualization tool, we have developed an automatic categorical axis reordering technique.

Because (unordered) categorical data by definition has no inherent ordering to the data values, it is possible to reorder the values per categorical axes without hindering interpretation of the values at that axes, which would not be the case with numeric data. Our technique reorders the categorical values based on the proportion of user-selected subpopulations with that value, causing categorical values with similar distributions to be located next to each other, enabling the user to more easily perceive similar categorical values with respect to the currently selected subpopulations.

Figure 1 provides an example of automatic categorical axis reordering. Figure 1a shows an overview of the radial coordinates visualization of 147 Primary Care Trusts (regional health care administrative bodies) in England. Twenty-six variables, such as cancer rates and socioeconomic factors, are represented by the radial axes, with each PCT represented by a curve connecting its value at each axis. Highlighted are the ONS_Area_Class_Group axis, a local regional classification (e.g. Manufacturing Towns and Coastal and Countryside) from the Office for National Statistics in the United Kingdom, and the lung_Combined_DSR axis, representing lung cancer rate in each PCT. Figure 1b shows a close up of the ONS_Area_Class_Group axis. In Figure 1c, the user has created two subpopulations with high (red) and low (blue) lung cancer rates. Reordering the categorical values for ONS_Area_Class_Group in Figures 1d and 1e make apparent that Industrial Hinterlands, Centres with Industry, Regional Centres, and Manufacturing Towns have the highest proportion of high lung cancer rate, whereas Prospering Smaller Towns, Prospering Southern England, London Suburbs, and Thriving London Periphery have the highest proportion of low lung cancer rate (New and Growing Towns at the bottom has no PCTs with high or low lung cancer rate). Categorical axis reordering thus aids in interpreting the relationship of other axes, including numeric axes, with categorical data. This is further explained in a paper presented to the 2014 Workshop on Visual Analytics, held in conjunction with the AMIA 2014 annual meeting.[3]

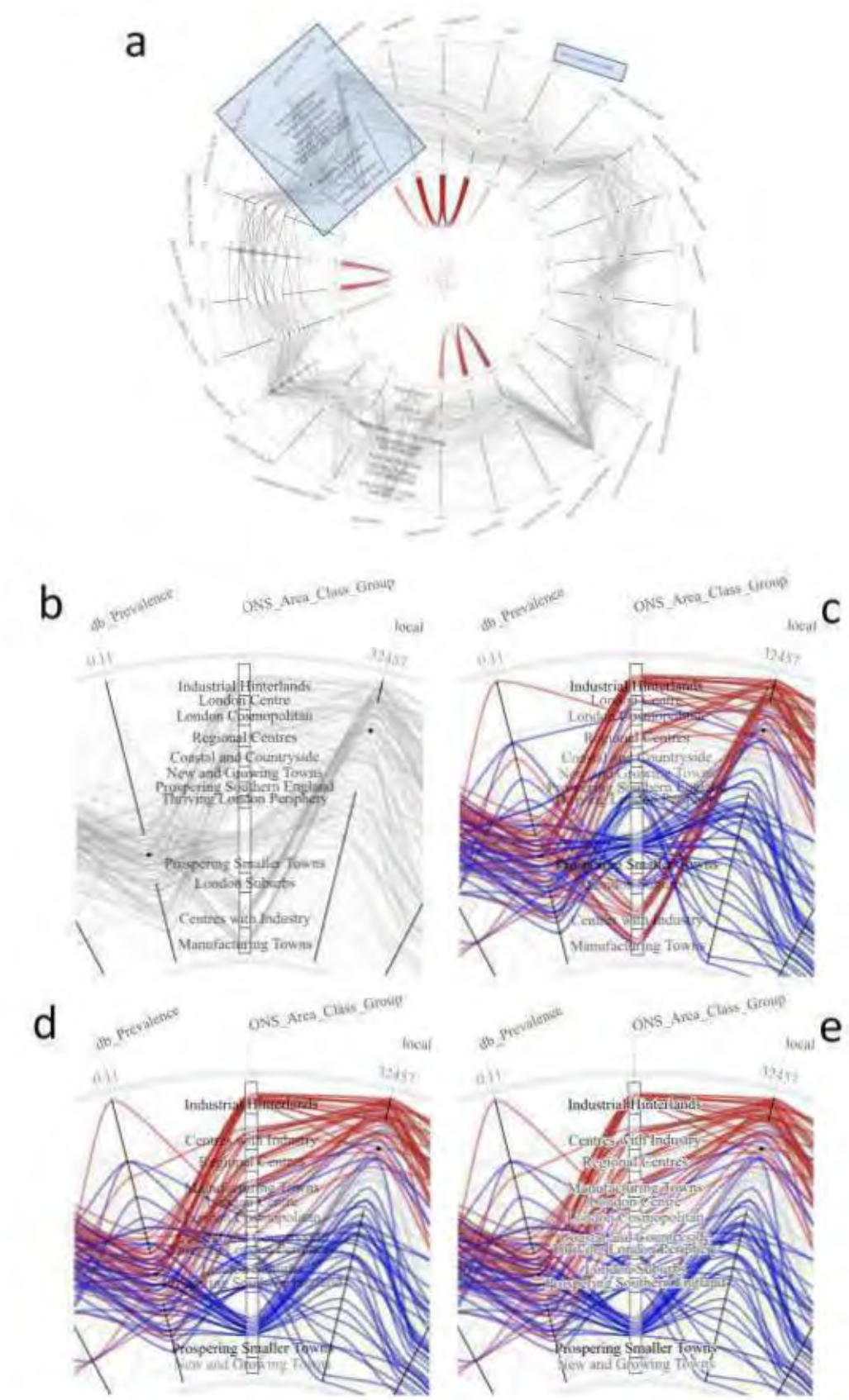


Figure 1: Radial coordinates with automatic value reordering of categorical axes

Temporal Visualization of Diabetes Data

We have developed a tool for the visualization of disease trajectories over time based on the parallel sets[4] visualization technique (Figure 2). For this technique, we used DEDUCE data from a query on patients with type 2 diabetes. Of interest in this work is that visualization of HgA1c values shows a large number of patients have lab values that return to normal in the last six months of life. We presented this work during the 2014 Workshop of Visual Analytics, held in conjunction with the AMIA 2014 annual meeting.[5-6]

In this visualization we align patients diagnosed with diabetes by death at the right, and visualize the paths of the HbA1c values of groups of patients going back in time from death, categorizing their lab values as Normal, Borderline, Controlled, and Uncontrolled. Data for the 535 patients is sampled every 6 months, and the users can control the sampling rate of the visualization above that. Figure 2a shows how much variability there can be in HbA1c values, and also shows a trend towards normalization at death (the green Normal section of the vertical axis at death is much larger than at the previous sample two years before death. The user can select any part of any path through the data to highlight that data and show the number of patients and percentage of the total represented by that path. Figure 2b highlights all patients with Normal HbA1c values at death, illustrating how many patients moved towards Normal at death.

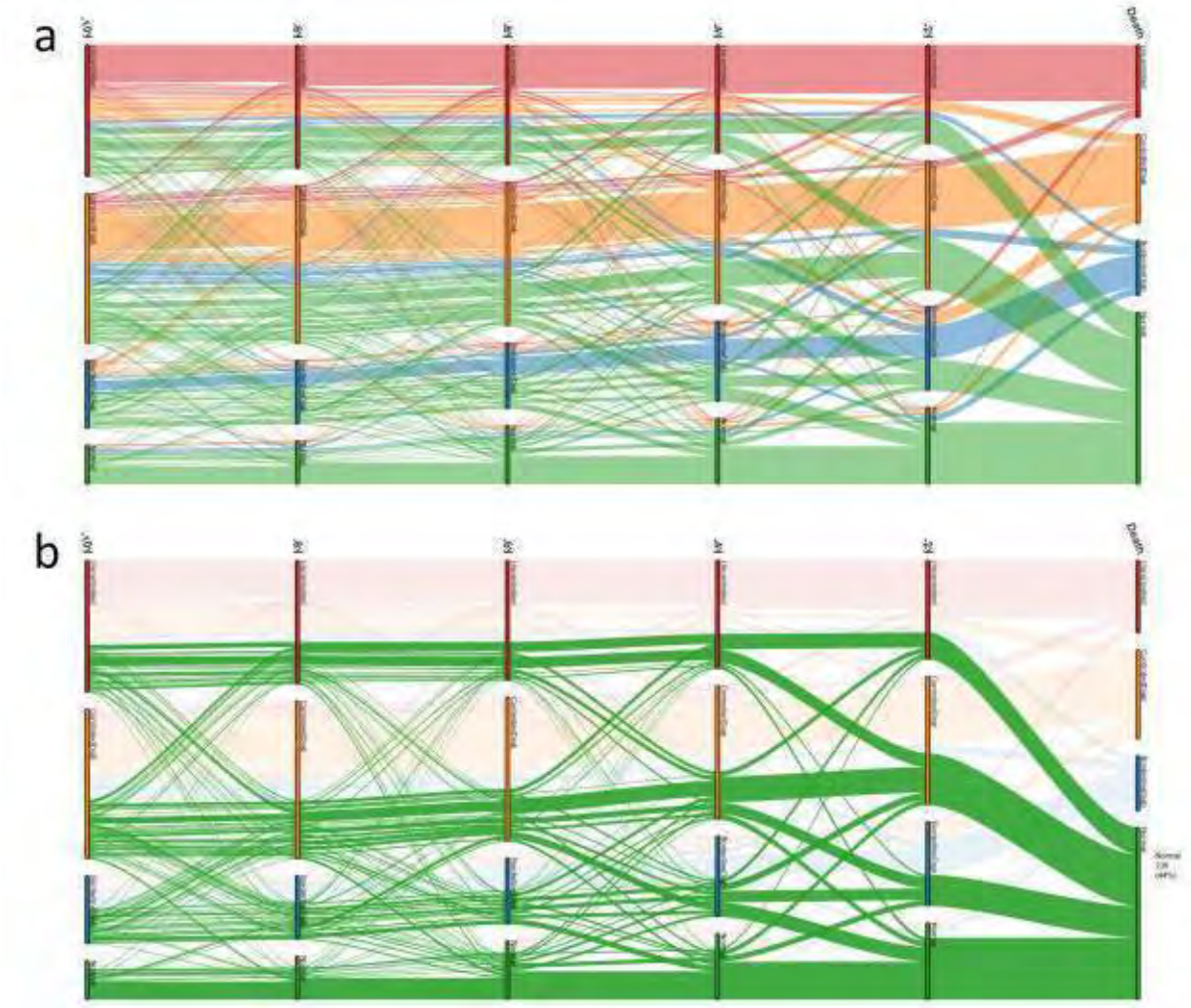


Figure 2: Parallel sets visualization of HbA1c values over time in diabetic patients.

We have added various controls to this visualization, such as the ability to show trajectories moving forward in time, Figure 3, where we compare the forward trajectories of patients who remained Uncontrolled (red) from -10 years to -4 years to those that remained Normal (green) from -10 years to -4 years.

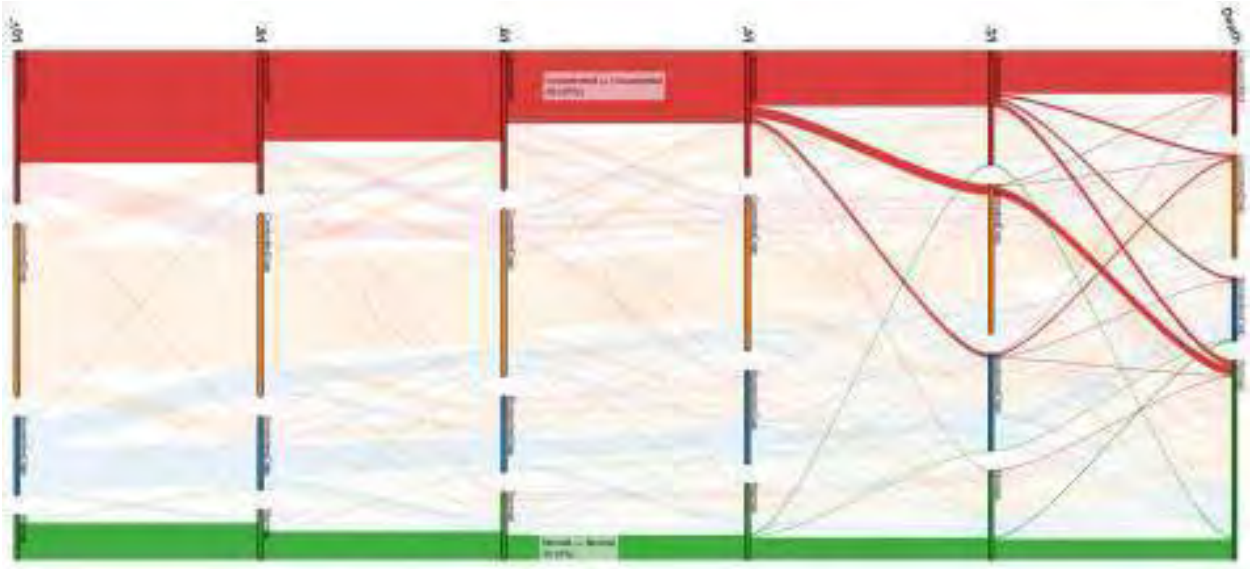


Figure 3: Comparing forward trajectories of *Uncontrolled* (red) and *Controlled* (green) starting 4 years before death.

We have also added the ability to incorporate non-temporal data, such as age, gender, and race, with our temporal visualization. Parallel sets is designed to enable the visualization of categorical data, so numeric data, such as age, must be transformed to an (ordered) categorical variable (e.g. decade). Categorical data can then be incorporated as additional vertical axes (Figure 4).

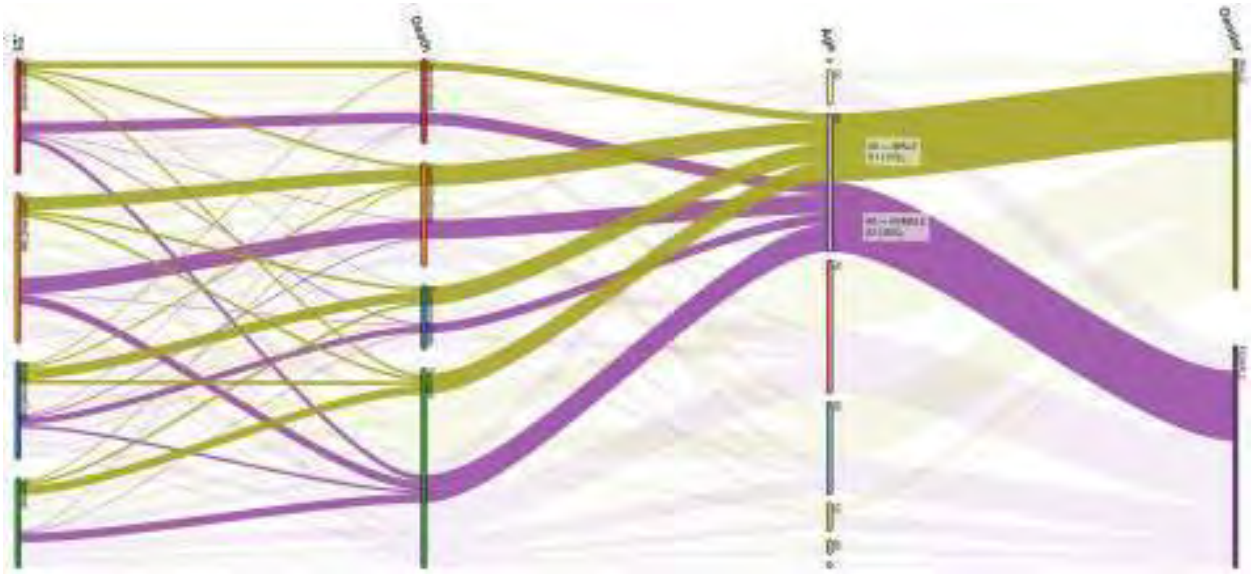


Figure 4: Adding numeric (*Age*) and categorical (*Gender*) axes. Here we compare the HbA1c trajectories between death and two years prior to death for males (yellow) and females (purple) in their 80s.

One of the advantages of the parallel sets temporal visualization is its ability to represent large numbers of patients, as paths are drawn for groups of patients with the same trajectory, with path width encoding relative number of patients with that path. A disadvantage of this visualization is the increased splitting of paths as more temporal samples are added, making it difficult to follow individual paths (Figure 5).

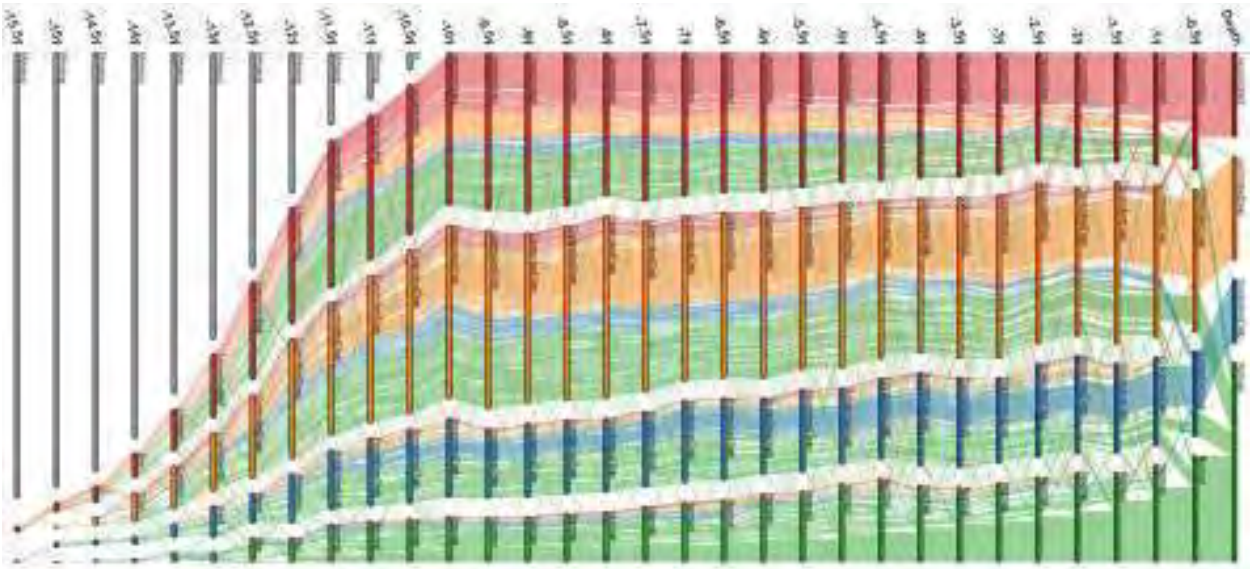


Figure 5: Parallel sets temporal visualization showing 15 years of data, sampled every 6 months.

We have therefore developed another temporal visualization technique, path maps, to enable enhanced perception of individual paths while still showing overall trends in the data. Our path map visualization is based on the heat map[7] technique, consisting of a 2D grid with each data cell assigned a color based on value and modifications specific for temporal data.

Figure 6 shows our path map visualization applied to the same data used for the parallel sets temporal visualization. Figure 6a shows the main path map visualization. Each row is an individual patient, and each column is a temporal sample point (every 6 months in this case), with data aligned by patient date of death on the right. The ragged left edge shows that patients had differing temporal ranges of data collection. HbA1c value categories are mapped color, with Normal as pale yellow through Uncontrolled as dark red. Figure 6b shows the distribution of each category across all temporal samples per patient, Figure 6c shows the distribution of each category across all patients per temporal sample, showing overall trends in the data, and Figure 6d shows the distribution of each category across all patients and temporal samples. A key feature of the path map visualization is how patients are sorted vertically. The user can click in any column to select that temporal sample, causing all patients to be sorted by their value at that sample. In Figure 6a, the user has selected a sample 5 years before death, highlighted by the black vertical lines. Black horizontal lines help the user visually segment the sample before and after the selected sample, with the pattern of colors in each segment showing the distribution of values. Various sorting methods can be used after sorting by the selected temporal sample to bring out different features of the data. Figure 6a uses a weighted average around the selected sample, whereas Figure 6e sorts first by the selected sample, then moving backwards from the

last sample. This method enables us to clearly see, for a given data value at a given time point, where those patients ended up. This data is emphasized by the thin colored lines within the highlighted column, which are colored by value at the last sample. For example, we can see that many more Uncontrolled patients ended up Normal than vice versa.

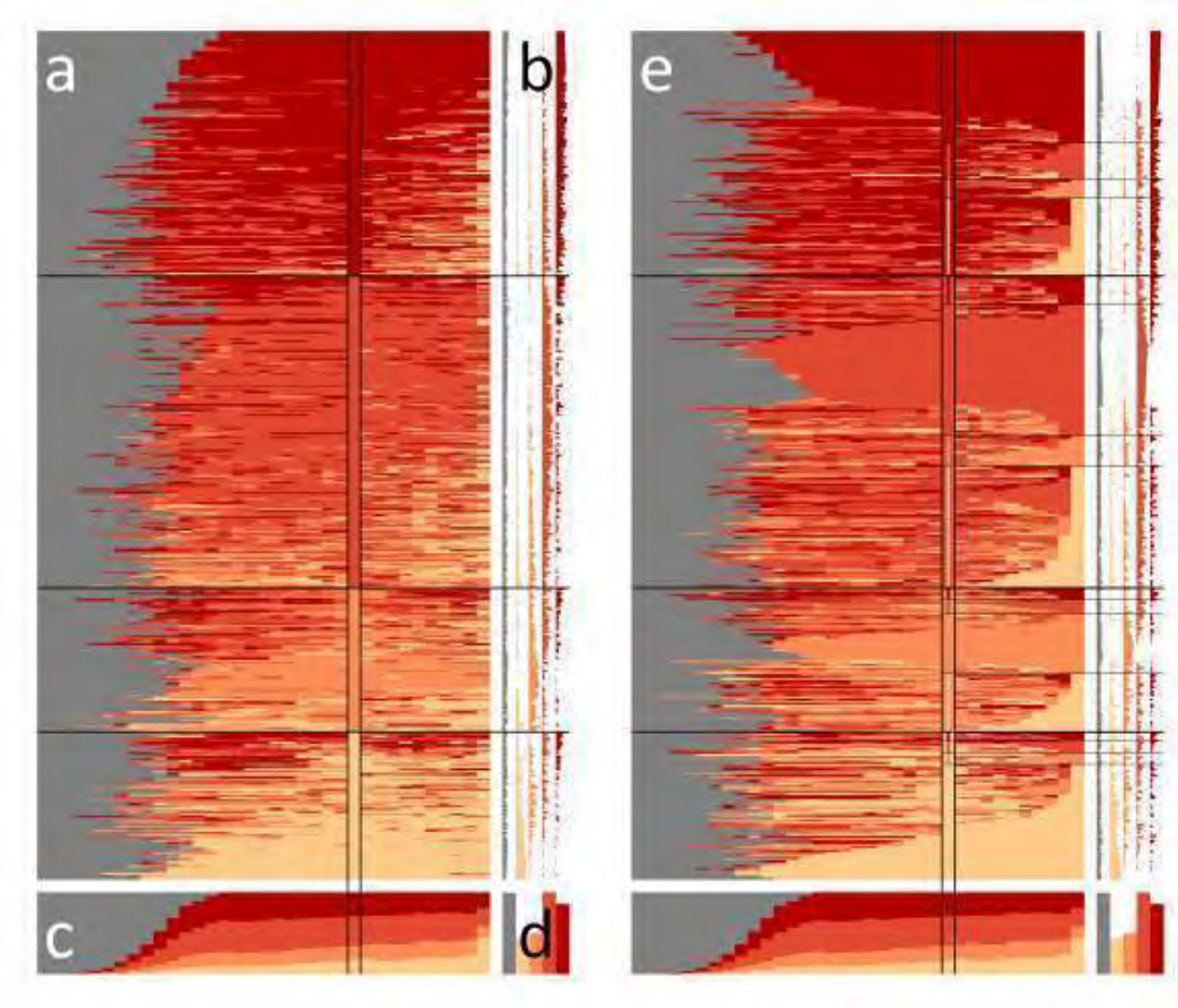


Figure 6: Path map temporal visualization tool

Although currently we do not incorporate other data, such as age, gender, or race, we anticipate doing so in two manners. One would include adding columns for each additional variable, similar to the technique used for the parallel sets temporal visualization, and the other would use a supplementary linked view, such as radial coordinates, in which patients could be selected in either view, and highlighted in the other.

Visualization of Free Text

Although not directly related to visualization of patient data, we have also experimented with visualizing free text data taken from the visualization in health care literature. Such visualizations could be useful with regards to patient notes. An abstract for a poster presentation of this work has been accepted for presentation at the AMIA Summits in San Francisco March 25, 1013.

Figure 7 shows a linked view visualization of documents, document clusters, topics, and terms, extracted from a textual analysis of the visualization in health care literature.

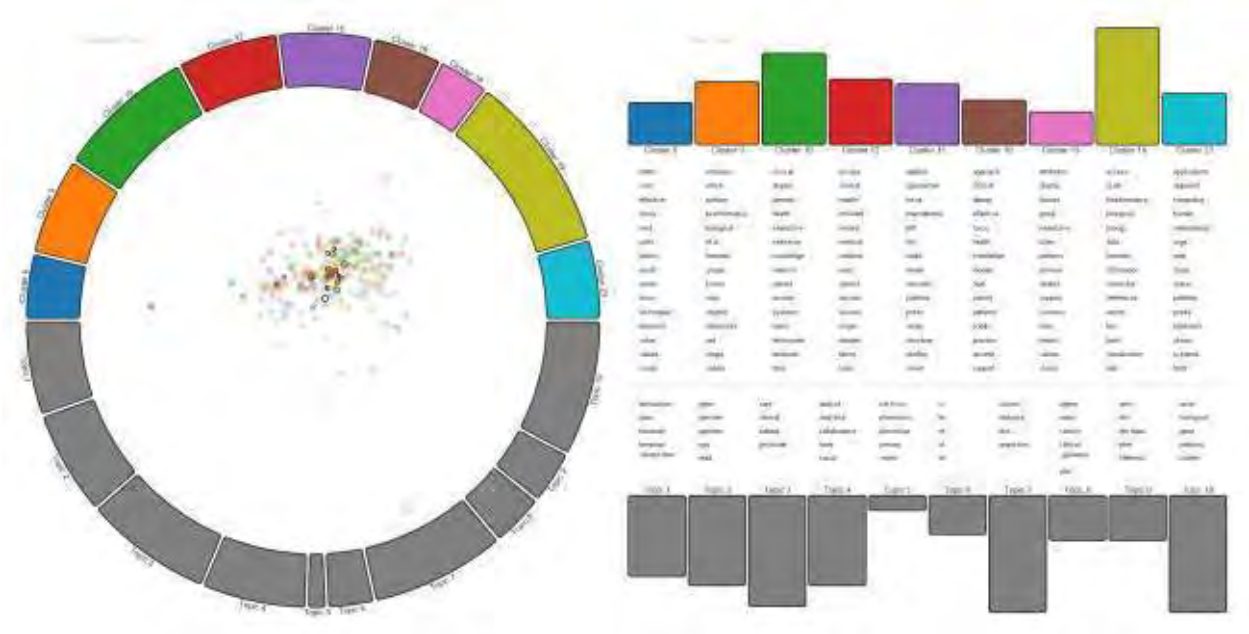


Figure 7: Visualization of visualization in health care literature

In Figure 8, the left view shows a scatter plot of documents based on single value decomposition, surrounded by colored clusters and grey topics. The right view shows terms associated with clusters and topics. Selecting any term highlights the documents, clusters, topics, and terms associated with that term.

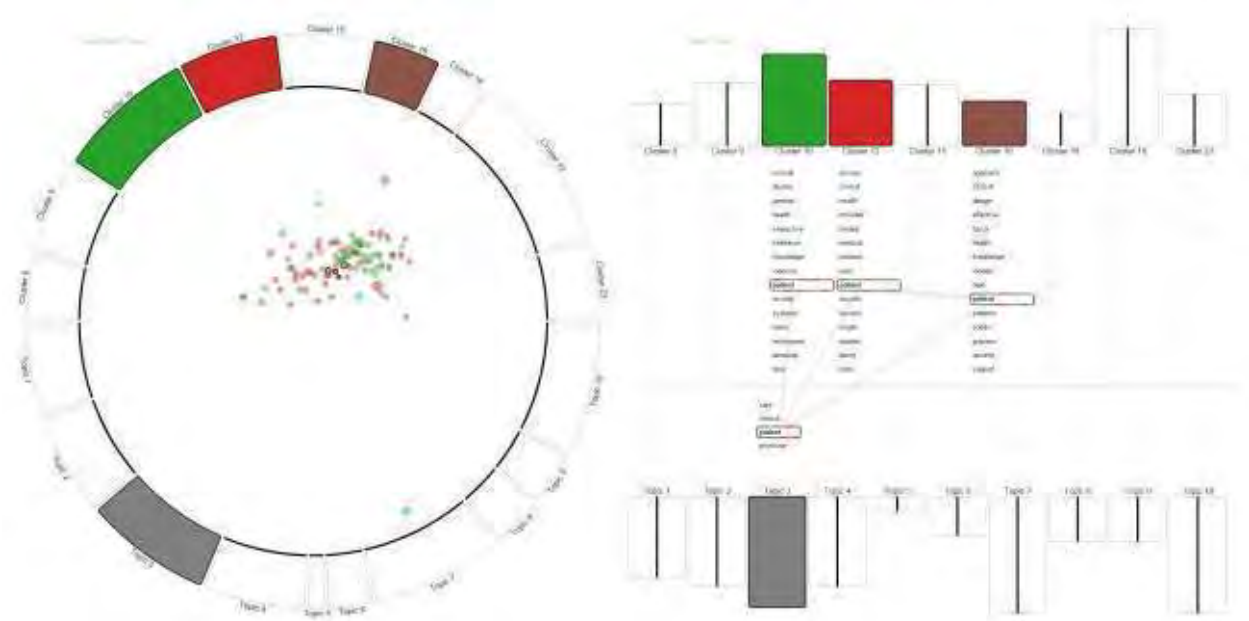


Figure 8: Selecting the term "patient" highlights all clusters and topics with that term, and all documents related to those clusters and topics.

Radial Coordinates

During the past year we have added features to our d3-based radial coordinates visualization tool, including correlation-based axis clustering, direct visualization of correlations via curved chords, and PCA-based scatterplots. This data is currently computed in R, and then loaded into the data visualization tool. Figure 9 shows these features applied to Primary Care Trust data from the NHS, with red chords connecting clusters of highly correlated axes, and a central scatterplot with linked selection to the radial coordinate curves.

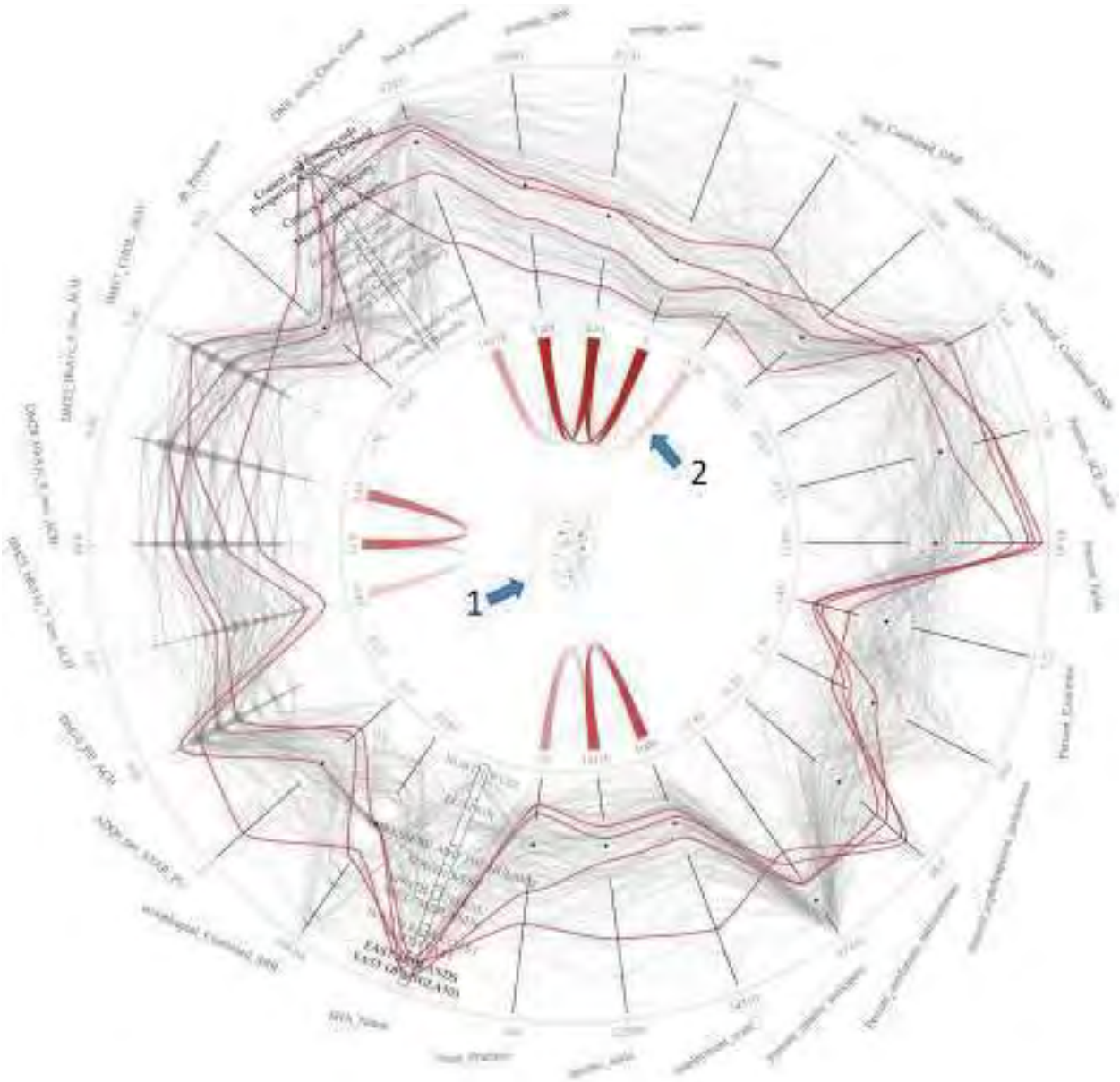


Figure 9: Radial coordinates visualization, showing PCA scatterplot (1) and correlation chords (2) in the central region.

We are in the process of applying our radial coordinates visualization to data sets resulting from DEDUCE data queries, for comparison with standard visualization approaches. Figure 10 shows an example ribbon rendering relating variables such as inpatient length of stay, age in years, admit location, and drug strength.

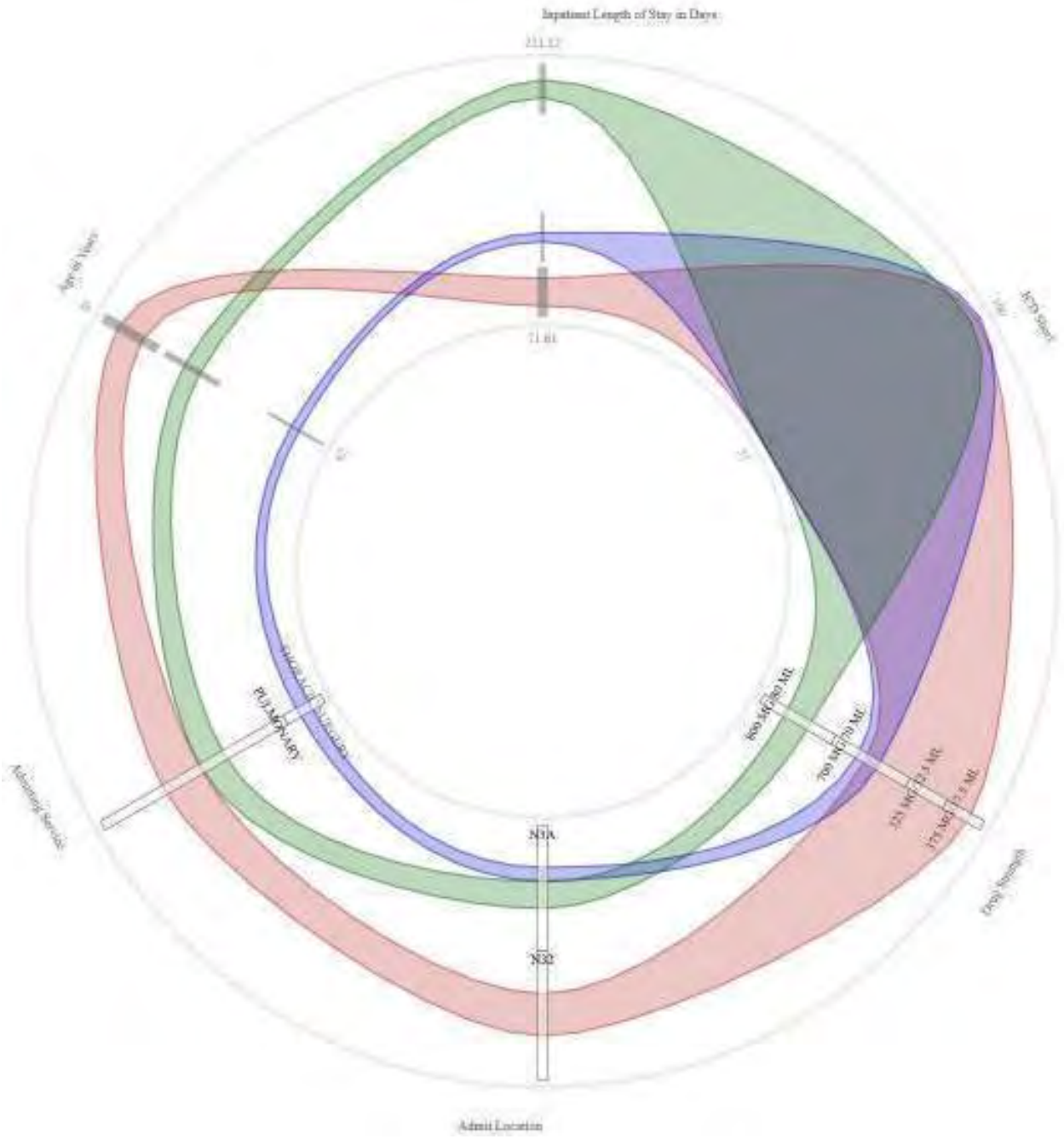


Figure 10: Radial coordinates visualization of data resulting from DEDUCE queries.

In the previous year we incorporated into our radial coordinates visualization tool axis distribution visualizations based on numeric type (continuous, discrete, categorical), represented by a stacked bar chart, histogram, and quartile plot respectively. These summary statistic visualizations enable visualization of the overall population distribution per axis. We also incorporated ribbon rendering to show distributions of user-selected subpopulations.

This year we have applied our radial coordinates visualization to data from the United Kingdom's National Health Service. The summary statistic visualization methods were shown to be useful in discovering a London suburb with a much higher esophageal cancer rate than other London suburbs (Figure 11a), and two health practices with much higher rates of a number of medical problems than other practices (Figure 11b). Our paper "Multivariate visualization of system-wide National Health Service data using radial coordinates" (See Appendix) provides more details.

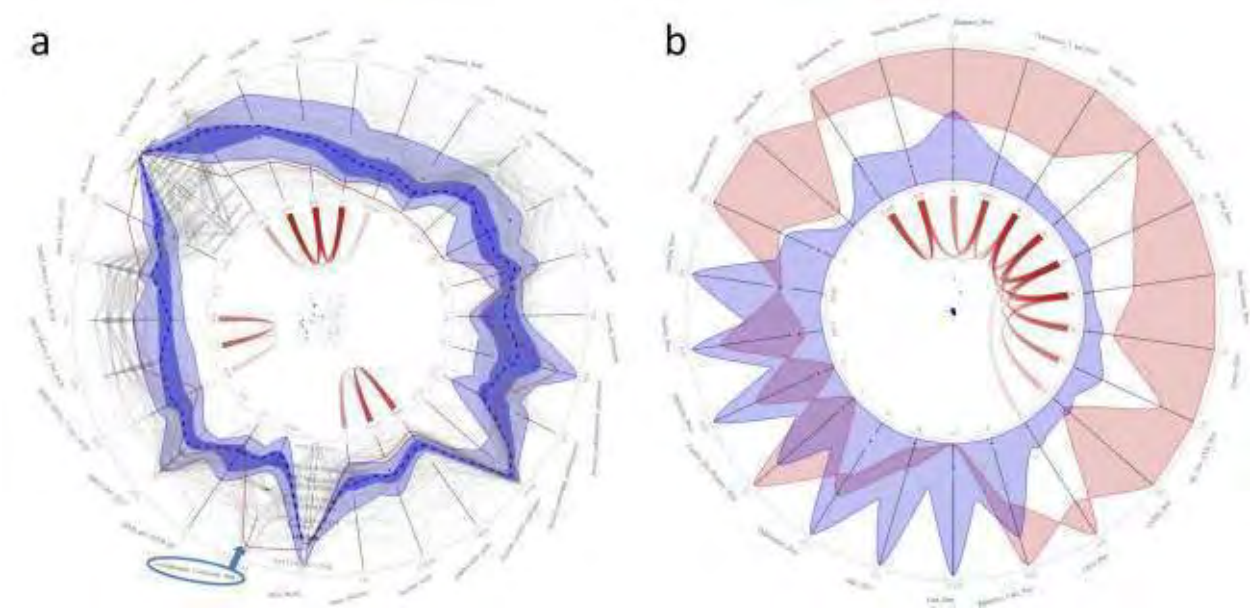


Figure 11: Radial coordinates visualization of Primary Care Trust (a) and practice-level (b) data in from the United Kingdom's National Health Service, using summary statistic visualization to discover interesting patterns in the data.

We have also begun work on connecting R to d3 via the Shiny library, and are able to calculate summary statistics in an R server on demand, and pass that data to a d3 visualization. We have not incorporated this ability fully into any of our visualizations, but anticipate use of this technique as we work with larger data sets.

2.8. Add supplemental data-dependent linked views of the data based on matrix of visualization techniques: Experiment with different layout patterns for parallel coordinates. STATUS of Milestone: Work is in process.

We have incorporated linked views into many of our visualization tools, such as the PCA scatterplot and radial view in our radial coordinates tool, and the smaller summary views linked to the main view in our path map tool. However, these tools all currently use linked views within the given visualization component. Our tool for visualizing the visualization in health care literature utilizes two separate components, a document view and a term view, linked together using D3's d3.dispatch object, a lightweight mechanism for loosely coupled components, such

that each component can listen to events from the other. Using this framework, we intend to more fully incorporate linked views such as linking selection between a temporal visualization, e.g. path map, and a multivariate visualization, e.g. radial coordinates.

We previously experimented with a "stair-step" layout of axes in an R-based prototype, but concluded that it was more confusing than a standard parallel layout. Our radial coordinates layout has proven useful for the visualization of large numbers of variables while maintaining a square aspect ratio. It also enables the use of the central region for supplemental visualizations, such as a PCA scatterplot of data entities, and chords showing correlations between axes (Figure 11 shows PCA scatterplot (a) and correlation cords (b) in the central region).

2.9. Complete testing visualization using PTSD. STATUS Work is ongoing.

Our experience with visualizing other health-related data sets, such as our radial coordinates multivariate visualizations of NHS data, and our parallel sets and path map temporal visualizations of diabetes data lays the groundwork in visualizing PTSD data. We will extend our radial coordinates visualization to work with larger data sets; as we add data elements to the visualization, the PTSD data may contains so much data that performance of the radial coordinates visualization will suffer. Our ribbon rendering technique provides a means for showing the distribution of user-defined subpopulations, so we plan on extending that technique such that it does not rely on each individual curve being drawn. The other major component that we will be adding is linking views to enable "drilling-down" into the data for each axis, for example, showing a parallel sets or path map visualization when selecting an axis with temporal data. Our experience designing linked views in our tool for visualizing the visualization in health care literature will prove useful here.

2.10. Future Work.

We will conduct interviews of users to evaluate the value of different visualizations and determine user responses to advanced visualizations of health care data. We will analyze the interview data using NVivo software.

We will finish adding additional variable to the diabetes visualizations and plan to submit a clinical manuscript to describe how visualization of lab values has discovered unexpected results: a large number of type 2 diabetic patients have normal HgA1c levels towards the end of life.

We will apply our radial coordinates visualization technique to PTSD data, incorporating linked views to enable drilling down into axes of interest, such as applying our temporal parallel sets or path map visualization to axes with temporal data. In addition, we will modify our radial coordinates visualization as necessary to handle any issues that arise with the PTSD data. These modifications may include changes to our ribbon rendering technique to enable the visualization of user-defined subpopulations without drawing curves for each data element, and using an R server via Shiny for calculating summary statistics for large data sets.

3. KEY RESEARCH ACCOMPLISHMENTS

The key research accomplishments emanating from this research to date are as follows.

- Completed classification of queries as those seeking information for research and those seeking information for quality improvement.
- Validated importance of interactive visualizations: force-directed network visualization provided information about DEDUCE queries that we were unable to find in a manual review.
- Found that visualization of HgA1c values of type 2 diabetic patients showed many more patients' blood levels normalized during the last 6 months of life, an unexpected result that we continue to investigate.
- Have incorporated axis clustering, correlation chords, and PCA scatterplot in d3-based radial coordinates visualization.
- Added automatic categorical axis value reordering to radial coordinates visualization.
- Developed d3-based parallel sets temporal visualization tool using diabetes data.
- Developed d3-based path map temporal visualization tool using diabetes data.
- Developed d3-based linked views for visualizing the visualization in health care literature.
- Completed proof-of-concept connecting R Shiny server to d3.

4. REPORTABLE OUTCOMES

The past year we report the following reportable outcomes.

4.1. JAMIA Publication

We submitted a manuscript to JAMIA's call for its Special Issue on Visual Analytics in Healthcare, The paper describes the systematic review of the literature we conducted last year to identify how visualization is used with health care data. Entitled Innovative Information Visualization of Electronic Health Record Data: A Systematic Review, it was accepted for publication. The Online First version for this Open Access article was published at the end of October; the journal is scheduled for release in March. (Copy in Appendix)

4.2. Abstracts and podium presentations

In its fifth year, the 2014 Workshop on Visual Analytics in Healthcare (VAHC 2014) was held in conjunction with the American Medical Informatics Association (AMIA) Annual Symposium in Washington, DC 15-19 Nov 2014. This day long Workshop provided an opportunity for participants to discuss visualization techniques, software applications, and datasets that are being used in various health care settings. We submitted two abstracts for peer review, both accepted as podium presentations: Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels and MultivariateVisualization of System-Wide National Health Service Data Using Radial Coordinates. (Copies in Appendix)

4.3 Abstract and demonstration

During VAHC 2014, a peer-review abstract was accepted as a demonstration for the afternoon session: Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c

Levels. During the demonstration, participants saw the data visualization techniques used for the diabetes mellitus abstract noted above. (Copy in Appendix)

4.4 Abstract and podium presentation: AMIA 2014 High School Scholars: Building New Paths to Biomedical Informatics Education

AMIA 2014 for the first time included a session for high school students who have participated in research with biomedical informatics programs in the U.S. In the summer 2014, a high school student worked with our research team and submitted an abstract and slides for peer review by a panel of AMIA members. She was one of six students selected to present the research she had been working on. Her presentation was entitled: Exploring Novel Visualizations of Survey Data from Users of Electronic Health Records. (Copy in Appendix)

4.5 Abstract for poster presentation: AMIA Summits 2015

We submitted an abstract for a poster presentation at the AMIA Summits 2015 in San Francisco. Entitled Visualization of the Healthcare Visualization Literature, it has been accepted for presentation on March 25.

5. CONCLUSION

The objective of our research is to explore interactive visualization of large sets of health data to provide better understanding of what is in the data. In the past year we have presented radial coordinates, a multivariate visualization technique based on parallel coordinates that incorporates features, such as per-axis population distribution visualizations based on data type (continuous, discrete, and categorical), direct visualization of correlations between variables, curve spreading for discrete and categorical data, visualization of summary statistics for user-selected subpopulations via ribbon rendering, and automatic reordering of categorical values based on user selection, driven by the needs of health-related data visualization. Using data from the BTS, we have illustrated the utility of the combinations of visualization techniques embodied in our radial coordinates tool.

Our hypothesis is that data visualization is more effective than traditional methods of data exploration, and that this type of visualization is highly dependent on the data and nature of the queries and what someone is trying to learn. We have been successful in visualizing HgA1c values of type 2 diabetes patients, with surprising results: the HgA1c values normalize during the last six months in a much larger population of the patients than expected. We will continue to add variables to this visualization and increase the number of patients to validate this finding.

We are excited about our results to data, and realize that much more can be accomplished by adding of other classes of data elements to visualize. For example, does environmental data increase the certainty of certain diagnoses? Does aggregation of patient data across sites of care increase the certainty value of diagnoses? What do genomic data and biomarkers add to the diagnosing certainty? What do patient reported outcomes contribute to diagnosing, determining the correct treatment, and caring for a patient? We propose that future research is needed to understand how to assign weighting factors and how to use this approach in a patient centric environment. We hope to continue our research when the current project is completed.

REFERENCES

1. Horvath MM, Rusincovitch SA, Brinson S, Shang HC, Evans S, Ferranti JM. Modular design, application architecture, and usage of a self-service model for enterprise data delivery: The Duke Enterprise Data Unified Content Explorer (DEDUCE). Journal of Biomedical Informatics. 2014:52: 231-242.
2. West VL, Borland D, Hammond, W. E. Innovative information visualization of electronic health record data: a systematic review. Journal of the American Medical Informatics Association. Published online First 21 Oct 2014. doi:10.1136/amiajnl-2014-002955.
3. Borland D, West VL, Hammond WE. Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 2014:53-58. https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf
4. Kosara R, Bendix F, Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics. July/August 2006;12(4):558-568.
5. McPeek Hinz E, Borland D, Shah H, West VL, Hammond WE. Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 2014:17-21. https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf
6. Shah H, Borland D, McPeek Hinz E, West VL, Hammond WE. Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 2014:53-58. https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf
7. Wilkinson, L. (1979), \Permuting a matrix to a simple pattern," in Proceedings of the Statistical Computing Section of the American Statistical Association , Washington, DC: The American Statistical Association, pp. 409{412.

.

APPENDIX

West, V. L., Borland, D., & Hammond, W. E. (2014). Innovative information visualization of electronic health record data: a systematic review. Journal of the American Medical Informatics Association, amiajnl-2014.

McPeek Hinz, E., Borland, D., Shah, H., West, V.L., & Hammond, W. E. (2014). Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 17-21. https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf.

Borland, D., West, V.L., & Hammond, W. E. (2014). Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 53-58. https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf.

Shah, H., Borland, D., McPeek Hinz, E., West, V.L., & Hammond, W. E. (2014). Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 53-58. https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf.

Ganapathiraju M. Exploring Novel Visualizations of Survey Data from Users of Electronic Health Records.

Downloaded from jamia.bmj.com on October 21, 2014 - Published by group.bmj.com

OPEN ACCESS

Innovative information visualization of electronic health record data: a systematic review

Vivian L West,[1] David Borland,[2] W Ed Hammond[1]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2014-002955).

[1]Duke Center for Health Informatics, Duke University, Durham, North Carolina, USA
[2]The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence to
Dr Vivian L West, Duke Center for Health Informatics, Duke University, 2424 Erwin Road, Suite 9002, Room 9021, Durham, NC 27705, USA; vivian.west@duke.edu

Received 5 May 2014
Revised 25 July 2014
Accepted 14 September 2014

ABSTRACT

Objective This study investigates the use of visualization techniques reported between 1996 and 2013 and evaluates innovative approaches to information visualization of electronic health record (EHR) data for knowledge discovery.

Methods An electronic literature search was conducted May–July 2013 using MEDLINE and Web of Knowledge, supplemented by citation searching, gray literature searching, and reference list reviews. General search terms were used to assure a comprehensive document search.

Results Beginning with 891 articles, the number of articles was reduced by eliminating 191 duplicates. A matrix was developed for categorizing all abstracts and to assist with determining those to be excluded for review. Eighteen articles were included in the final analysis.

Discussion Several visualization techniques have been extensively researched. The most mature system is LifeLines and its applications as LifeLines2, EventFlow, and LifeFlow. Initially, research focused on records from a single patient and visualization of the complex data related to one patient. Since 2010, the techniques under investigation are for use with large numbers of patient records and events. Most are linear and allow interaction through scaling and zooming to resize. Color, density, and filter techniques are commonly used for visualization.

Conclusions With the burgeoning increase in the amount of electronic healthcare data, the potential for knowledge discovery is significant if data are managed in innovative and effective ways. We identify challenges discovered by previous EHR visualization research, which will help researchers who seek to design and improve visualization techniques.

BACKGROUND AND SIGNIFICANCE

In 2004 a presidential executive order, 'Electronic Health Records (EHRs) for All Americans', laid out tenets to improve the quality and efficiency of healthcare, with one goal being accessible EHRs for most Americans within 10 years.[1 2] In September 2009, years of research and policy work culminated in the Health Information Technology for Economic and Clinical Health Act (HITECH Act) allocating $19.2 billion in incentives to increase the use of EHRs by hospitals and health delivery practices. The latest report from the Centers for Medicare and Medicaid Services (CMS) found that approximately 80% of eligible hospitals and over 50% of eligible professionals had received incentive payments from CMS for adopting EHRs.[3]

With the burgeoning amount of electronic data, the potential for knowledge discovery is significant if the large amounts of data are managed in innovative and effective ways. This review investigates data visualization techniques reported in the healthcare literature between 1996 and 2013, aiming to evaluate innovation in approaches to information visualization of EHR data for knowledge discovery.

Historical background

The graphical visualization of data dates back to the later part of the 18th century when William Playfair is credited with the first use of line graphs, pie charts, and bar graphs. Playfair, an engineer and economist, considered charts and graphs the most effective way to communicate information about data.[4] In 1786, he published *The statistical breviary; shewing, on a principle entirely new, the resources of every state and kingdom in Europe; illustrated with stained copper plate charts, representing the physical powers of each distinct nation with ease and perspicuity*, stating that by graphically representing data, the reader can best understand and retain the information.[5]

A widely recognized visualization is a two-dimensional graph using time, temperature, and geography showing Napoleon's march on Russia in 1812. This linear graph, published by Charles Minard in 1861, shows the movement of Napoleon's army across Russia to Moscow and back to Europe (figure 1).

The horizontal axis of the graph is marked with temperatures below freezing as they returned. The width of the line depicting the army is scaled, illustrating the decline in the number of men returning from war as the temperature decreased, which can be easily compared to the size of the army as they set out in 1812. Tufte and Graves-Morris point out that Minard's innovative graph relies on six variables: size (of army), latitude and longitude (where the army was), direction (that army was moving), location (at certain dates), and temperature (where the army was).[6]

One of the first effective means of using medical data to generate knowledge was developed in 1858 when Florence Nightingale used a polar-area diagram (also called a coxcomb chart) to demonstrate the relationship between sanitary conditions and soldiers' deaths compared to death from battlefield wounds (figure 2).[7 8]

Since then, standardized charts and graphs have been used for specific types of healthcare data to quickly determine the need for appropriate interventions. For example, graphing vital signs data can quickly identify a rise or fall in physiological data, indicating the need for an intervention and demonstrating the effectiveness of the intervention; and

To cite: West VL, Borland D, Hammond WE. *J Am Med Inform Assoc* Published Online First: [*please include* Day Month Year] doi:10.1136/amiajnl-2014-002955

Copyright 2014 by American Medical Informatics Association.

Downloaded from jamia.bmj.com on October 21, 2014 - Published by group.bmj.com

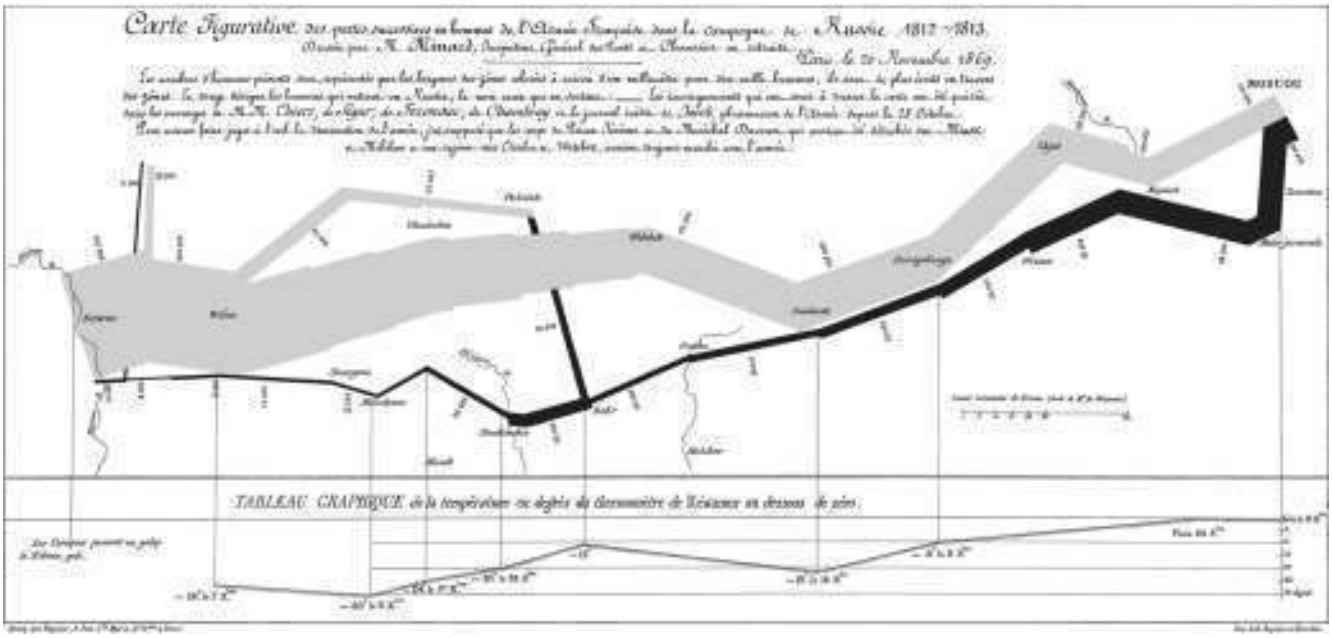


Figure 1 Translation: 'Figurative chart of the successive losses in men by the French army in the Russian campaign 1812–1813. Drawn up by Mr. Minard, inspector-general of bridges and roads (retired). Paris, 20 November 1869. The number of men present is symbolized by the broadness of the colored zones at a rate of 1 mm for ten thousand men; furthermore, those numbers are written across the zones. The red [note: gray band here] signifies the men who entered Russia, the black those who got out of it. The data used to draw up this chart were found in the works of Messrs. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished journal of Jacob, pharmacist of the French army since 28 October. To better represent the diminution of the army, I've pretended that the army corps of Prince Jerôme and of Marshall Davousz which were detached at Minsk and Mobilow and rejoined the main force at Orscha and Witebsk, had always marched together with the army.' Public domain (U.S.) image via Wikimedia Commons. Available at http://commons.wikimedia.org/wiki/File:Minard.png (accessed 21 July 2014).

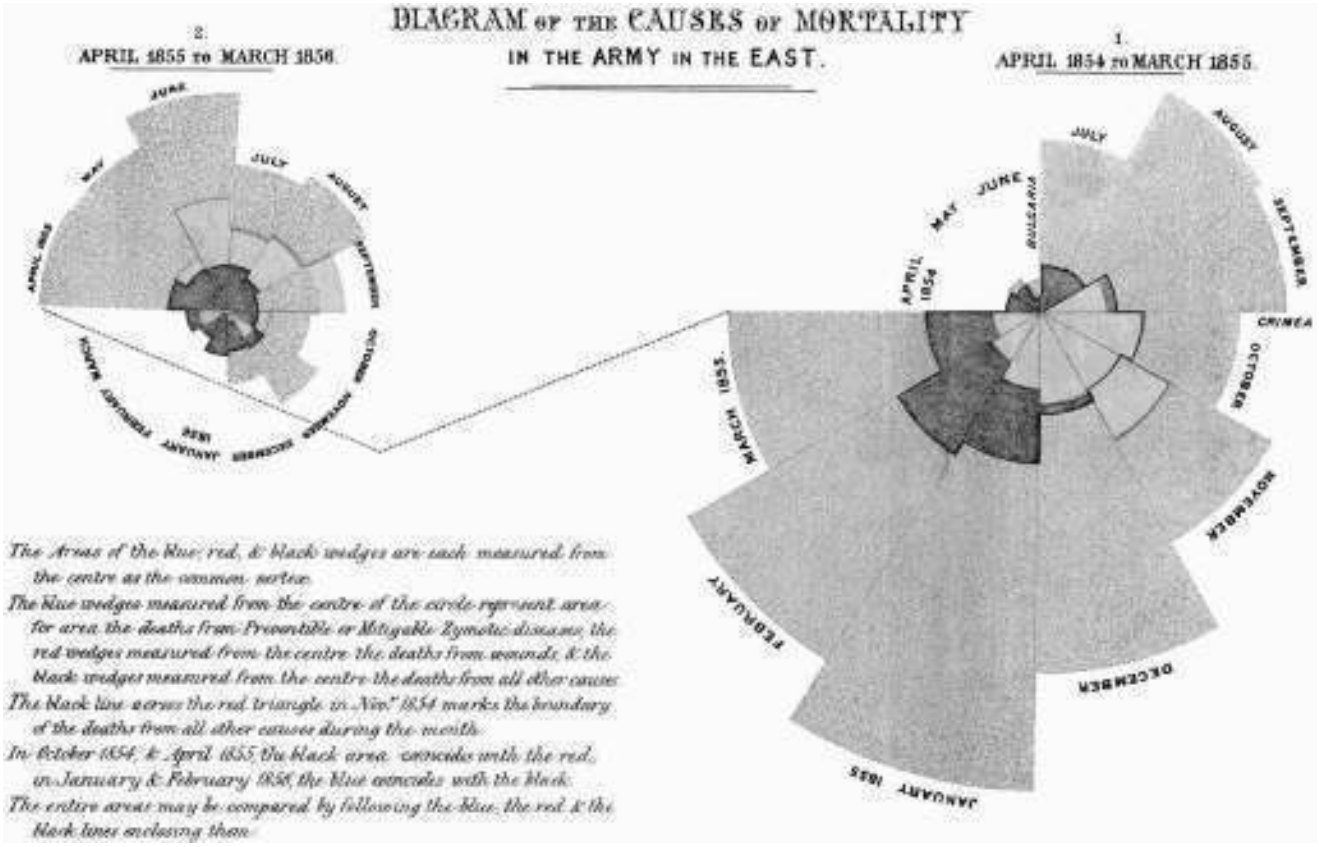


Figure 2 Florence Nightingale's coxcomb chart representing causes of death each month between April 1854 and March 1856 during the Crimean War. The *large outer gray bands* represent deaths attributed to lack of sanitation in the wards, the *lighter gray middle bands* to death from wounds during the war, and the *darkest inner bands* to other causes. Public domain (U.S.) image via Wikimedia Commons. Available at http://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg (accessed 21 July 2014).

Downloaded from jamia.bmj.com on October 21, 2014 - Published by group.bmj.com

Fishbone diagrams are commonly used graphic representations of laboratory results.

A plethora of scales, shapes, and colors have been used with both small and large datasets rendered as visual diagrams such as bar charts, line graphs, scatterplots, and pie charts to reveal patterns leading to knowledge discovery. Industries such as finance, accounting, and the petroleum industry routinely use information visualization, defined as 'interactive, visual representations of abstract data to amplify cognition',[9] using innovative approaches that account for both the volume and complexity of their data. In the healthcare field, however, applications of advanced visualization techniques to large and complex EHR datasets are limited.

Data in healthcare

In 1994, Powsner and Tufte[10] proposed summarizing patient status with test results and treatment data plotted on a graph. This was one of the earliest examples of using several diverse datasets in medical records to visualize information. Also in the 1990s, Plaisant *et al*[11] developed LifeLines as a means to visualize patient summaries using several different graphical attributes, for example colors and lines depicting a patient's discrete events. Furthermore, Shahar and Cheng[12 13] developed Knowledge-based Navigation of Abstractions for Visualization and Explanation (KNAVE) as a means to explore time-oriented, semantically-related concepts.

Clinical records by nature contain longitudinal data of patient visits over time, with records of changing problems, medications, treatments, and responses related to evolving health status. Graphs are routinely used to illustrate data in a way that comparisons, trends, and associations can be easily understood. In healthcare studies, the use of graphs with time as the horizontal axis to display various types of data has been increasing, and several well established visualization tools have been developed using temporal data, with LifeLines/LifeLines2[11 14–16] and KNAVE/KNAVE-II/VISITORS[17–20] the most widely reported. When querying 'longitudinal studies' in PubMed, 7071 publications were found in 1983, with the number consistently rising through the following 30 years to 45 821 studies published in 2013.

Longitudinal data from EHRs displayed through innovative visualization techniques has tremendous potential for discovering useful information in the data. Until health record data became widely available electronically, however, there was little emphasis on using such large and complex datasets. We argue that EHR data is actually a new kind of data that requires new visualization techniques beyond graphs and charts to accommodate the size of the dataset and explore the contents of the data.

Exploring EHR data with visualization techniques other than tables, graphs, and charts is a nascent approach to understanding the information in EHRs. A comprehensive monograph by Rind *et al*[21] focuses on a survey of visualization systems and criteria typically used by designers of systems. A book by Combi *et al*,[22] and two book chapters[23 24] also describe several of the visualization systems reported in the literature. We report our results from a systematic review that describes how innovative visualizations are being used with large and complex EHR data as a means to present or 'discover' information without specific hypotheses.

Objectives

The aim of this review is to investigate the visualization techniques that have been used with EHR data and answer the following questions:

- What is the prevalence of the use of information visualization with EHR data?
- Are techniques being used for knowledge discovery with an entire EHR dataset?
- What has been learned from research on visualization of EHR data?

METHODS

We conducted a systematic literature review following the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement.[25] Our review was limited to articles published between 1996 and 2013. We began with 1996, the year one of the largest healthcare systems in the USA, the Veterans Health Administration, first mandated the use of EHRs.[26] The Health Insurance Privacy and Portability Act (HIPPA) of 1996 was also enacted to provide security of individually identifiable health information, with a consensus that EHRs would be the most effective way to assure compliance with HIPPA. It is also the year when the first study using visualization with complex data (medical records histories and associated longitudinal data) was published by Plaisant *et al*.[11] This time interval enables us to construct the historical timeline for the use of information visualization in healthcare, particularly as data have become more common electronically due to the legislative requirement for conversion to EHRs.

An electronic literature search was conducted in May–July 2013 using MEDLINE, the most frequently used reference database in healthcare, and Web of Knowledge. This was supplemented using citation searching and gray literature searching. Reference lists from highly relevant articles were also reviewed to find additional articles. Broad keywords were used to assure a comprehensive document search (see table 1).

Inclusion and exclusion criteria

Articles had to include the use of EHR data using innovative visualization techniques, or describe developing techniques that would be applied to EHR data. We define EHR data as data in electronic clinical records that contain clinical information (eg, demographics, problems, treatments, procedures, medications, labs, images, providers) collected over time that can be shared among all authorized care providers. We define innovative visualizations as visualizations other than standard graphs traditionally used for displaying healthcare information (such as bar charts, pie charts, or line graphs) that use complex data, which we define as data with multiple types of variables and many data points, resulting in an exceptionally large amount of data, such as that in an entire EHR. We were interested in any innovative visualization technique for vast amounts of information that might be the foundation for an interactive system; therefore, although interaction is a key characteristic of information visualization, we included articles describing static visual

Table 1 Search terms used in search

Keyword	Boolean	Additional keywords
Information visualization		
Information visualization	AND	Health data, electronic health record, electronic medical record
Visualization	AND	Big data, clinical data, health data, health care data, healthcare data, electronic health record, electronic medical record

Downloaded from jamia.bmj.com on October 21, 2014 - Published by group.bmj.com

representations of large amounts of EHR data in addition to interactive visualizations.

Articles were excluded if they related to animals or plants, were position papers describing the need for visualizing data or ideas for techniques in visualization, or did not describe specific techniques used for the visualization or have figures showing the results from visualization. The literature is replete with articles on visualization in genetics, syndromic surveillance, and geospatial environmentally aware data, which we also did not include in our review because we were focused on clinical EHR data as defined above. There were many articles on the technical details related to visualization techniques, which did not fit our target for studies explaining how clinical data is used in visualization; these were also excluded.

Article selection and analysis

The authors, title, journal, year of publication, and abstract for each article were collected in an Excel spreadsheet. To identify key themes for matrix analysis, the first 50 abstracts and titles were reviewed; 11 themes were identified. These themes were then added to the spreadsheet to form a matrix for reviewing and categorizing all abstracts and to assist with determining which should be excluded.

After reviewing all abstracts and eliminating those categorized with exclusion criteria or lacking inclusion criteria, full articles of the remaining were read for eligibility. Our primary interest in conducting the study was to understand what innovative information visualization techniques in healthcare have been reported using EHR data since 1996. The review is not a meta-analysis and does not include a statistical analysis. The objective of the study was to investigate the prevalence of information visualization techniques used with EHR data, therefore we did not conduct a risk of bias assessment.

RESULTS

A total of 847 references were retrieved from our initial search of electronic databases, specifically MEDLINE (PubMed and PMC) and Web of Knowledge. A search of the gray literature and hand-searching references from articles yielded an additional 44 papers. All abstracts were reviewed, with duplicates removed (n=191). We then excluded 666 articles because the visualizations discussed were diagnostic, did not relate to EHR data, focused on animals or plants, used genomics data, discussed geospatial data or syndromic surveillance, were position papers suggesting the need for visualization or describing a potential visualization technique, or were primarily discussions of the technical details of visualization.

The full text of each of the remaining 34 articles was then read; 16 of these articles were excluded (table 2 lists for reasons for exclusion). Results of the screening process in the analysis are noted in the flow diagram in figure 3.

Eighteen articles were included in the qualitative synthesis. The online supplementary table S3 summarizes those included.[11 14–20 27–36]

The studies reviewed describe prototypes in various stages of development. Four of the articles describe LifeLines, the most advanced application, with its continued revisions and application in various populations. First described in 1996 by Plaisant *et al*,[11] LifeLines was developed as a prototype using data from the Maryland Department of Juvenile Justice to provide a visual overview of one juvenile's record on a single screen. LifeLines, using electronic health data, provides a timeline of a single patient's temporal events; time is represented on the horizontal axis, and events (problems, allergies, diagnoses, complaints, labs, imaging, medications, immunizations, communication) are listed vertically.

Table 2 Articles excluded from analysis

Reason for exclusion	No.	Explanation
Article not applicable	9	Articles are medical guidelines, no visualization is described, or articles describe process
Visualization not applicable	1	Does not use EHR data
Geospatial information	1	
Genetics	1	
Position paper	3	Ideas for visualization
Technical	1	
Total	16	

EHR, electronic health record.

LifeLines evolved to LifeLines2 and the use of multiple patient records. LifeLines2 research found that users want to see both numerical and categorical data, and that the ability to drill down into details when looking at patient records is an important feature.[16] Several other visualization techniques have been developed by this team using multiple records, for example LifeFlow,[31] developed for use with millions of patient records visualized on a single page that allows the user to see trends and evaluate quality of care. Using LifeFlow, new users can easily explore the data to understand patterns and trends at a high level.

Four articles[17–20] describe a second innovative visualization called VISITORS, or Visualization of Time-Oriented Records. VISITORS is based on earlier work of Shahar and colleagues, whose research conceptualized clinical data (eg, multiple measures of temperature over time) summarized into abstractions (in this case, fever). This was KNAVE[12]; KNAVE-II is a later enhancement.[18] Like LifeLines, VISITORS applies what researchers learned in earlier applications for a single record to develop an application that accommodates diverse temporal data from multiple records. Usability testing found the system feasible for exploring longitudinal data for quality or clinical results. The interface used with VISITORS was deemed to need simplification, in spite of the short time it took for users to learn how to use the system.

One article that we might have excluded used relatively simple linear graphs to illustrate the correlation of abstract concepts with laboratory values.[29] The data used in the analysis are from the 3 million patient EHRs for New York Presbyterian Hospital, promising complexity in the data. Both factors, EHR data and complex data, are inclusion criteria; therefore, the study was included in the final analysis. Seven laboratory tests and sign-out notes used primarily by residents to assist overnight staff caring for inpatients were abstracted. From the sign-out notes, 30 clinical concepts were identified using pattern matching, and then correlated with normalized lab values graphed on a timeline. The research showed the value of using time in the correlation, and the value of using aggregated data from many records versus a single record. It also demonstrates how temporal patterns can be visually found in EHR data using pattern matching and temporal interpolation.

A different approach is proposed by Joshi and Szolovits[34] using a radial starburst to show the complexity of data represented over a 100-dimensional space. The complexity is reduced by using machine learning to group similar clusters of patients

Downloaded from jamia.bmj.com on October 21, 2014 - Published by group.bmj.com

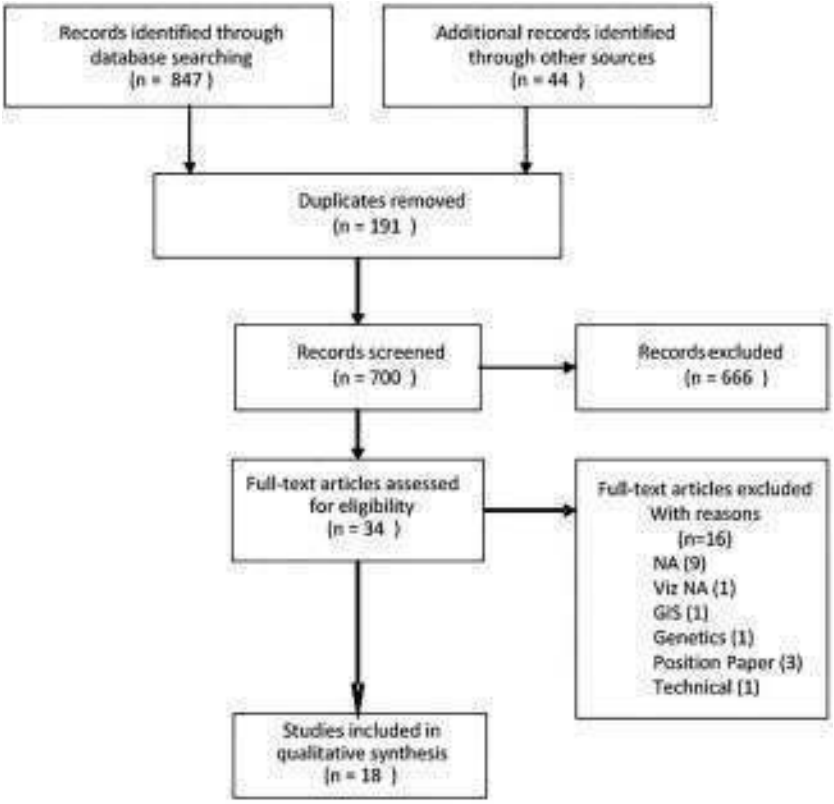


Figure 3 Flow of information through the different phases of systematic review. Adapted from the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) group.[25]

characterized by eight physiological foci to allow a user to look at one patient and evaluate the severity of that patient's condition. This is an example of using a very large set of data, or the EHR 'big data' as a clinical decision making tool. Although it is a static representation, Joshi and Szolovits provide a visual representation with an interesting presentation of complex data that has potential as a foundation for an interactive system with interactions such as filtering, selection, or brushing (highlighting a subset of data).

Gotz *et al*[30] developed Dynamic Icons, or DICON, as a visualization technique for exploring clusters of similar patients. By applying algorithms to EHR data, they found clusters of patients similar to the target patient. The user can interactively explore the clusters represented as icons on a treemap. They found this visual representation, a unique approach to visualization of healthcare data, required time for users to understand. Once users understood the design concept, however, the interface provided functionality for rapidly analyzing the data using icons that could be easily controlled.

Gotz and Wongsuphasawat[32 33] designed Outflow as a means to look at disease progression paths based on the assumption that the onset of a particular disease symptom applies perpetually, with common disease states among patients and transitions between the states. Outflow allows users to look at a visual display consisting of multiple events, their sequences, and outcomes to quickly analyze the event sequences in order and accurately identify factors most closely correlated with specific pathways.

Wang *et al* report using LifeLines2 and sentinel event data for subject recruitment to clinical trials.[19] They found using alignment, ranking, and filtering functions reduced user interaction time when working with sentinel events. Its use for subject recruitment was found to be questionable, however. Data in medical records can be somewhat uncertain, making the timeline inaccurate. For example, a patient with a long-standing diagnosis of asthma who visits a care provider for shortness of breath may be coded as first being diagnosed with asthma on that visit, even though the diagnosis of asthma was made previously. If a clinical trial includes patients diagnosed with asthma within a certain time range, the patient would be excluded in recruitment.

Fifteen studies address the use of temporal data.[11 14–20 27 29 31–35] Most articles describe interactive visualizations. All but two articles focus on use of the visualizations for clinical decision support. The two visualizations not used for decision support suggest use for quality assurance and improvement.[25 28]

Most studies that included an evaluation of the visualization described the training of the user and training time. One study reported training time of 6 min for its visualization, which used radial displays with a body map in the center of the radius and the relevant physiological parameters highlighted on the body map.[36] This was the shortest training time reported; the longest was a half hour.[30]

Although several ways to visualize EHR data are described, it was difficult to discern if the data as described were actually real-time data, or retrospective data or databases with predetermined datasets. Some of the articles describe systems for data visualization, for example, LifeLines2 and VISITORS. Others use visualization techniques such as sequential displays,[31 36] treemaps,[28 30] radial displays,[34 36 38] or icicle trees.[31]

DISCUSSION

Although most studies recognize the importance of the growing amount of clinical data, we found few innovative EHR visualization techniques that lend themselves to the large amount of data available electronically. Prior to 2010, seven publications we reviewed employed different and innovative visualization techniques with healthcare data; three of those describe LifeLines and

Downloaded from jamia.bmj.com on October 21, 2014 - Published by group.bmj.com

three describe KNAVE-II/VISITORS. With the HITECH Act in 2009, national interest in EHRs was high, with increasing interest in knowledge that might be discovered by using visualization techniques applied to EHR data. Three studies on visualization of electronic health data were reported each year in 2010 and 2011, and four in 2012. Data from 2013 are not inclusive since our review was conducted in May–July 2013 (figure 4).

Several themes are common: the type of data accessible to the user, meaningfulness of visualizing large amounts of data, usability, and training time. Challenges from research to date can be broadly categorized into four areas identified by Keim *et al*[37] in other domains using very large, complex datasets: data (quality, size, diversity), users (needs, skills), design (intertwining both in a system that provides an easy way to visually explore and analyze results), and technology (tools, infrastructure).

Research on EHR visualization provides some important lessons on challenges that need to be addressed:

- The amount of EHR data and its display is a challenge; the more data, the more difficult it is to see and identify meaningful patterns in visualizations. Using tools such as zoom, pan, and filter reduces some of the clutter, but the purpose of the visualization will affect the use of such tools. If researchers are to use entire datasets from EHRs to discover information within the data, it will be necessary to develop better ways to manage the massive amounts of data.
- The size and complexity of EHR data is a challenge. Color, density, and filtering techniques are commonly used to distinguish variables or temporal events. Although scaling and zooming have been used to resize data, none of the reported techniques in the studies we reviewed discuss applicability to an entire EHR dataset and the potential for knowledge discovery in this very large composite dataset.
- The ability to use temporal data in visualizing aggregate data from EHRs is important to users.
- Researchers need to be cognizant of the many variables that can lead to uncertain data in EHRs; uncertain data can distort temporal events.
- EHR data are complicated by missing values, inaccurate data entry, and mixed data types that must be considered in developing visualization techniques.
- Presenting a great deal of information in a single screen shot where the user can interactively explore the information is an important design feature.
- Users want to see both categorical and numerical data when interactively exploring the data, and they like to look at the detail in the record. This is challenging with visualizing an extremely large amount of data in an EHR, but important for user satisfaction.
- A normalization scheme is needed for aggregated numerical data.
- The time it takes to learn the system is an important consideration that is complicated by the complexity of the data using visualizations that are different from those most clinicians and researchers are used to seeing, such as charts and graphs.
- Training time to understand and effectively use the visualization for its intended purpose should be considered when developing visualization techniques. Training is usually the user's first introduction to visualization. The complexity of the visualization and ways to navigate the display will increase training time if it is not easy to explain or demonstrate the functionality of the visualization.

Aigner *et al* have identified similar challenges working with temporal data, which is inherent in EHR data: the complexity, quality, diversity, and uncertainty of data; the interfaces and roles of the users; and evaluation of quality and effectiveness of the design.[38] The interest and challenges in data analysis with 'visual presentation and interaction technologies' that can be used with very large and complex datasets is universal.[39] The ability to explore and gain a deeper understanding of the value of 'big data' will encourage adoption of visualization techniques in healthcare. Research focused on these challenges is needed if we are to fully utilize EHR data for knowledge discovery.

Limitations

Although there are numerous articles published by Plaisant *et al* and Shahar and Klimov that are related to the techniques incorporated in their specific visualizations (LifeLines/LifeLines2/LifeFlow/EventFlow and KNAVE/KNAVE-II/ VISITORS), our review was limited to those articles that were the primary publications describing the innovative visualization technique and its application to electronic health data. By restricting our review to a narrow segment of this literature, we may have inadvertently eliminated meaningful details from our review.

Our search terms were intentionally broad; we eliminated articles whose abstracts indicated the articles were more technical in nature, and we eliminated articles whose focus was on geospatial representation. We may have obtained different results had more specific terms been used.

Finally, there are books and book chapters that deal with visualization of healthcare data. These types of publications are not

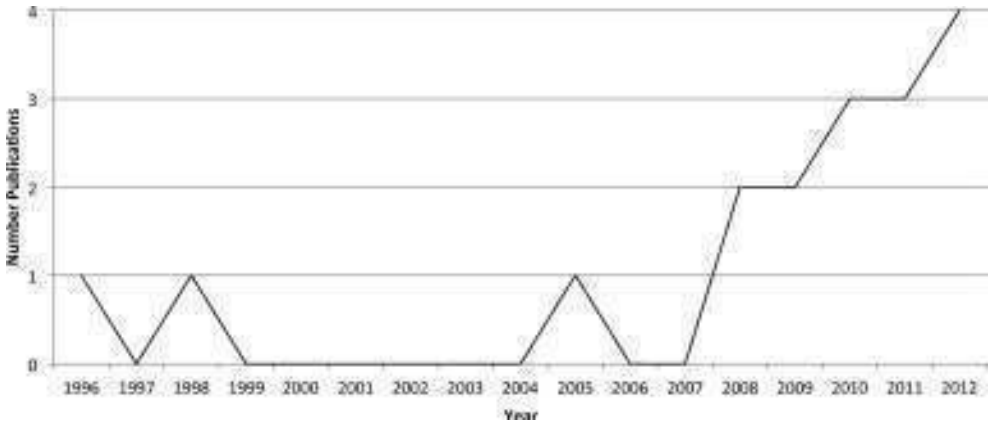


Figure 4 Number of publications included in review.

Downloaded from jamia.bmj.com on October 21, 2014 - Published by group.bmj.com

included in our review, but may contain information relevant to this review.

CONCLUSIONS

This study was conducted to determine the prevalence of the use of information visualization for EHR data, what techniques have been used, and what research has taught us to date. Although there is increasing interest in visualization of electronic healthcare data, few techniques have been found to effectively and efficiently display the large and complex data in EHRs.

The new buzzword in healthcare is 'big data', often used in conjunction with data analysis. Most studies have found that visualization of EHR data requires techniques that will handle not only 'big data', but the temporal complexity of constantly changing variables found within EHR data. Disciplines such as computer science, engineering, and genetics have developed visualizations to improve presentation, analysis, and understanding of data. The healthcare provider community has not yet taken advantage of these methods or significantly explored the use of new visualization techniques to accelerate the use and understanding of EHR data. We have identified important findings reported in the literature that can help guide future research needed to explore, refine, and retest visualization techniques. Only then will stakeholders begin to take advantage of the wealth of knowledge within EHR data.

Acknowledgements The authors wish to acknowledge Rene' Hart, Staff Assistant at the Duke Center for Health Informatics, for her help in organizing the literature review.

Contributors VLW conducted the literature search. All authors contributed to the analysis and text. VLW and DB provided the figures. All edited, reviewed, and approved the final version.

Funding This work was supported by the US Army Medical Research and Materiel Command (USAMRMC) grant number W81XWH-13-1-0061.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

REFERENCES

1 Bush GW; Office of the Press Secretary, the White House. Executive Order: Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator. Press release, April 27, 2004. http://www.whitehouse.gov/news/releases/2004/04/print/20040427-4.html2004 (accessed 23 Jul 2013).
2 Bush GW. State of the Union Address, Promoting Innovation and Competitiveness, President Bush's Technology Agenda. 2004.
3 Data Show Electronic Health Records Empower Patients and Equip Doctors. Press release, July 17, 2013. http://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-Releases/2013-Press-Releases-Items/2013-07-17.html (accessed 23 Jul 2013).
4 Spence I. William Playfair and the psychology of graphs. *American Statistical Association JSM Proceedings*; 2006:2426–36.
5 Playfair W. The statistical breviary; shewing, on a principle entirely new, the resources of every state and kingdom in Europe; illustrated with stained copper plate charts, representing the physical powers of each distinct nation with ease and perspicuity. To which is added, a similar exhibition of the ruling powers of Hindoostan. London: J Wallis, 1801.
6 Tufte ER, Graves-Morris PR. *The visual display of quantitative information.* Vol 2. Cheshire, CT: Graphics Press, 1983.
7 Nightingale F. Notes on matters affecting the health, efficiency, and hospital administration of the British Army. Founded chiefly on the experience of the late War. Presented by request to the secretary of state for War. Privately printed for Miss Nightingale, Harrison and Sons, 1858.
8 Lienharg J. The Engines of Our Ingenuity, Episode 1712: Nightingale's Graph. Podcasts between 1988–2002. 1988–2002. http://www.uh.edu/engines/epi1712.htm (accessed 21 Jul 2013).
9 Card SK, Mackinlay JD, Shneiderman B, eds. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.
10 Powsner S, Tufte E. Graphical summary of patient status. *Lancet* 1994;344:386–98.
11 Plaisant C, Milash B, Rose A, *et al*. LifeLines: visualizing personal histories. *SIGCHI Conference on Human Factors in Computing Systems Proceedings*; 1996:221–7.
12 Shahar Y, Cheng C. Intelligent visualization and exploration of time-oriented clinical data. Systems Sciences, HICSS-32, 32nd Annual Hawaii International Conference Proceedings 1999 (Volume:Track4).
13 Shahar Y, Cheng C. Model-based visualization of temporal abstractions. *Comput Intell* 2000;16:279–306.
14 Plaisant C, Mushlin R, Snyder A, *et al*. LifeLines: using visualization to enhance navigation and analysis of patient records. *AMIA Symposium Proceedings*; 1998:76–80.
15 Wang TD, Plaisant C, Quinn AJ, *et al*. Aligning temporal data by sentinel events: discovering patterns in electronic health records. *CHI '08 SIGCHI Conference on Human Factors in Computing Systems Proceedings* 2008:457–66.
16 Wang TD, Wongsuphasawat K, Plaisant C, *et al*. Visual information seeking in multiple electronic health records: design recommendations and a process model. *1st ACM International Health Informatics Symposium Proceedings*; 2010:46–55.
17 Klimov D, Shahar Y. A framework for intelligent visualization of multiple time-oriented medical records. *AMIA Annu Symp Proc* 2005;2005:405–9.
18 Martins SB, Shahar Y, Goren-Bar D, *et al*. Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data. *Artif Intell Med* 2008;43:17–34.
19 Klimov D, Shahar Y, Taieb-Maimon M. Intelligent interactive visual exploration of temporal associations among multiple time-oriented patient records. *Methods Inf Med* 2009;48:254–62.
20 Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artif Intell Med* 2010;49:11–31.
21 Rind A, Wang TD, Aigner W, *et al*. Interactive information visualization to explore and query electronic health records: a systematic review. *Foundations Trends Hum-Comput Interact* 2013;5:207–98.
22 Combi C, Keravnou-Papailiou E, Shahar Y. Temporal information systems in medicine. Springer, 2010.
23 Aigner W, Kaiser K, Miksch S. Visualization techniques to support authoring, execution, and maintenance of clinical guidelines. Computer-based Medical Guidelines and Protocols: A Primer and Current Trends 2008;139:140–59.
24 Lesselroth BJ, Pieczkiewicz DS. Data visualization strategies for the electronic health record. Nova Science Publishers, Inc., 2011.
25 Moher D, Liberati A, Tetzlaff J, *et al*.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
26 Anderson M. Lessons learned from the Veterans Health Administration. Healthc Pap, 5, 30–37. 2005. https://www.thecsiac.com/sites/default/files/files/Clinger%20Cohen%20(1996).pdf (accessed 15 Apr 2013).
27 Bashyam V, Hsu W, Watt E, *et al*. Problem-centric organization and visualization of patient imaging and clinical data. *Radiographics* 2009;29:331–43.
28 Willison B. Advancing Meaningful Use: Simplifying Complex Clinical Metrics Through Visual Representation. Parsons Institute for Information Mapping (PIIM) Research 2010.
29 Hripcsak G, Albers DJ, Perotte A. Exploiting time in electronic health record correlations. *J Am Med Inform Assoc* 2011;18(Suppl 1):i109–115.
30 Gotz D, Sun J, Cao N, *et al*. Visual cluster analysis in support of clinical decision intelligence. *AMIA Annu Symp Proc* 2011;2011:481–90.
31 Wongsuphasawat K, Guerra Gómez JA, Plaisant C, *et al*. LifeFlow: visualizing an overview of event sequences. *SIGCHI Conference on Human Factors in Computing Systems Proceedings* 2011:1747–56.
32 Gotz D, Wongsuphasawat K. Interactive intervention analysis. *AMIA Annual Symposium Proceedings*; 2011:274.
33 Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Trans Vis Comput Graph* 2012;18:2659–68.
34 Joshi R, Szolovits P. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. *AMIA Annu Symp Proc* 2012;2012:1276–1283.
35 Stubbs B, Kale DC, Das A. Sim•TwentyFive: an interactive visualization system for data-driven decision support. *AMIA Annu Symp Proc* 2012;2012:981–900.
36 Zhang Z, Wang B, Ahmed F, *et al*. The five W's for information visualization with application to healthcare informatics. *IEEE Trans Vis Comput Graph* 2013;19:1895–1910.
37 Keim DA, Kohlhammer J, Ellis G, *et al*., eds. *Mastering the information age-solving problems with visual analytics*. Florian Mansmann, 2010.
38 Aigner W, Federico P, Gschwandtner T, *et al*. Challenges of Time-oriented Data in Visual Analytics for Healthcare. *IEEE VisWeek Workshop on Visual Analytics in Healthcare*; 2012.
39 Thomas JJ, Cook KA. A visual analytics agenda. *IEEE Comput Graph Appl* 2006;26:10–3.

Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels

Eugenia McPeek Hinz[1], David Borland[2], Hina Shah[3], Vivian L. West[3], W. Ed Hammond[3]
[1]Duke Health Technology Solutions, Duke University; [2]RENCI, The University of North Carolina at Chapel Hill;
[3]Duke Center for Health Informatics, Duke University

Abstract

Diabetes mellitus is a chronic long-term disease requiring consistent medical treatment to achieve glucose control and prevent complications. Time of diabetes diagnosis can be variable and delayed years beyond disease onset. The spectrum of glycemic trajectories for a general population over an entire diabetes disease course is not well defined. Aligning disease course on death enables coherent data visualization. Our temporal visualization tool uses a parallel-sets inspired technique that illustrates the complicated and varied trajectories of hemoglobin A1c levels for a general diabetic population. A consistent glucose normalization trend for the majority of patients is seen over the course of their disease, especially in the six months prior to death. This tool permits discovery of population-level Hemoglobin A1c trends not otherwise evident without disease phase synchronization. These findings warrant further investigation and clinical correlation. Visualizations such as this could potentially be applied to other chronic diseases and spur further discoveries.

Introduction

Diabetes mellitus is a chronic disease that affects millions worldwide, resulting in numerous cardiovascular and renal complications, and subsequently is a major cause of death. Age of onset, duration of diabetes, and poor glycemic control are well-defined risk factors for the development of complications associated with increased mortality in persons with diabetes mellitus.[1] To decrease the development of complications associated with diabetes, tightly controlled glucose is the standard of care.[2] Notably some large prospective trials have found either worse outcomes or lack of benefit for some patients at high risk for complications under tight treatment control regimens.[3,4] Hemoglobin A1c (HbA1c), a marker of glucose control over the two to three months preceding the test, is a validated predictor of diabetes-related complications.[2] Using HbA1c to understand trajectories and temporal patterns of glycemic control over an entire diabetes disease course could be an important factor in improving treatment and reducing overall complications.

Data visualization techniques offer opportunities to explore large datasets and identify clinical patterns that might otherwise not be obvious. In this study we present a cohort of patients with diabetes (via ICD9 codes) from Duke University's data warehouse, visualizing their HbA1c levels over time, aligned by death, to explore trends of glycemic control. To the best of our knowledge, temporal visualization of glycemic control for a diabetic population standardized on death has not previously been presented. Our visualization groups HbA1c values into ordered categories of glycemic control (Normal, Borderline, Controlled, and Uncontrolled), utilizing a method based on parallel sets[5] and Sankey diagrams[6] to view temporal patterns in HbA1c values. We incorporate a number of features to facilitate interactive data exploration, such as viewing the progression of values either forwards or backwards in time, the ability to change the temporal sampling and range of the data being viewed, highlighting of multiple subpopulations, coloring based on the category along each path in the data or at the beginning/end of each path, and the incorporation of demographic data, such as gender.

Related Work

Analysis of diabetes indicators

A reduction in HbA1c levels lowers the risk of diabetes-related complications and mortality, especially for patients earlier in their disease course.[7] Counterintuitively, intensive treatment of glucose to reach near-normal levels for patients already experiencing diabetes-related complications has failed to lower all-cause mortality.[3] While large cross-sectional studies of populations such as the National Health and Nutrition Examination Survey find a temporal trend toward improving glycemic control over time, less well-established is the temporal trajectory of glycemic control for diabetic patients in general.[8] The only other work the authors are aware of looking specifically at glycemic control trajectories for a large diabetic cohort followed patients prospectively to the end point of death.[9] The study correlated initial glucose control to outcome of death, but did not report specifically on the population glucose trajectories.

Visualization methods

Our visualization tool is based on parallel sets[5] and Sankey diagrams.[6] Parallel sets were originally developed for visualizing relationships in multivariate categorical data, whereas Sankey diagrams, introduced by M. H. P. R. Sankey, are typically used for describing the flow of quantities such as energy, material, or cost. The original parallel sets user interface enables user-defined classification definitions, statistical analysis information, and various sorting methods. Parallel sets combines the concepts of parallel coordinates[10] and mosaic plots[11], enabling an aggregation of data points within visualization elements, as opposed to showing each individual element, which is typical of parallel coordinates. Multiple systems aggregate data points for summary.[5,12-14] For example, EventFlow enables the search and visualization of interval data, such as periods of medication treatment, to examine the order of sequences of events in the data.[12] OutFlow facilitates analyses of temporal event data in the form of pathways with relevant statistics.[14] All of these visualization tools look at event occurrences and their order, without placing these events on time axes. Our diabetes visualization uses the parallel sets paradigm, with each axis representing a temporal sample of HbA1c levels instead of a separate variable, similar to von Landesberger et al.[13] Although our current dataset is relatively small (121 patients), we chose a parallel sets representation in part due to its ability to aggregate many data points. The visual complexity is bounded by the number of axes and categories per axis, not by the number of data points, making it suitable for the exploration of larger datasets in the future. This representation also easily incorporates additional non-temporal variables, such as demographic data.

Methods

Data extraction and preprocessing

Data from Duke University's data warehouse were extracted using DEDUCE, an on-line query tool developed at Duke to assist researchers in human subjects research and departments seeking quality improvement data.[15] Beginning with over 4.4 million patients, we first queried by 23 IDC9 codes for diabetes mellitus, with and without complications. The query was further refined by querying on patient death indicator and laboratory tests for glycosylated hemoglobin (HbA1c), and finally by including only patients prescribed anti-hyperglycemics. This search returned data from 208 patients. From this cohort of 208, we eliminated four that did not have a year of death recorded, one whose date of death was documented but continued to have laboratory results recorded after that date, and 82 who did not have at least 10 years of HbA1c laboratory values. Our final cohort includes data from 121 patients.

We average HbA1c values, given as a percentage of total hemoglobin, over 6 month time intervals. In the case of missing HbA1c values within a 6 month period we first attempt to impute an HbA1c value from the average glucose (AG) values over that period of time, via the formula $HbA1c = (AG + 46.7) / 28.7$.[16] If no glucose values exist in that time period, the previous HbA1c value (measured or imputed) is carried forward. HbA1c values are then classified into four categories based on the severity of diabetes: Normal < 5.7, Borderline $[5.7, 6.5)$, Controlled $[6.5, 8)$, and Uncontrolled ≥ 8.

The sampled data is time-aligned by the death event for each patient. The visual representation of diabetes progression propagates backwards in time initially. Time is represented as number of years before death.

Visualization

Our visualization tool was developed using the D3 Javascript library.[17] The aim of this visualization is to investigate temporal trajectories of HbA1c levels for a large cohort of diabetes patients over a number of years prior to death. Since parallel sets is effective for showing relations between categories using aggregated frequencies of paths through categories at each dimension, it is a reasonable choice for showing HbA1c summary trajectories. The visualization tool shows a total of five categories: four representing glycemic control, and one optional Missing category for patients with data greater than 10 years before death. . Each vertical axis is a time step. The user can choose the frequency of these time steps, with a minimum sampling frequency of six months. The user can also select the maximum number of years before death.

The death event axis is placed at the right with all other time steps moving backwards in time to the left (Figure 1). Each vertical axis is split into the four HgA1c categories (Normal in green, Borderline in blue, Controlled in orange, and Uncontrolled in red), and a Missing category in grey prior to 10 years before death. The height of each axis category represents the proportion of the patients in that category at that point in time. Paths moving between axes recursively split moving backwards from death to show the trajectories of similar groups of patients. The visualization can show trends either starting at the death event i.e. going backwards in time (dividing

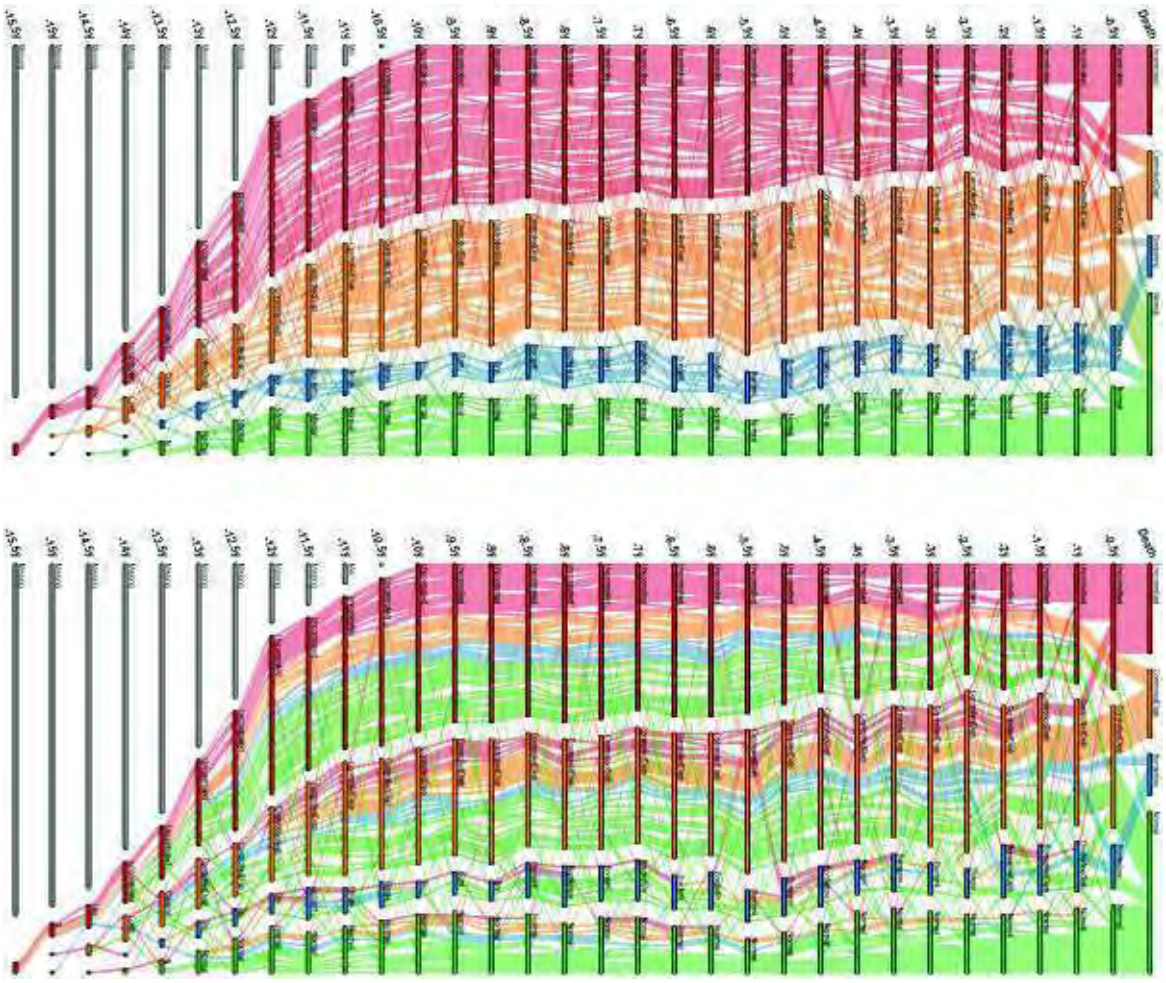


Figure 1. Diabetes progression overview visualizations. The top image colors paths by the current HbA1c at each time step, which is useful for emphasizing overall temporal trends. The bottom images colors paths by the HbA1c level at death, showing at each time step where each path will end.

recursively right to left), or starting at the last year in the visualization, i.e. going forward in time (recursive division from left to right). Going backwards and coloring by death shows at any time point the relationships between patients in a given category and their categories at death, while going forward in time shows the relationship between patients in a given category and their categories at a user-defined earlier point in time (Figure 2). Following Shneiderman's Mantra[18] of first overviewing and then filtering, the user can highlight one or more groups of patients by clicking on categories or trajectories to highlight the behavior of that group of patients going backward and forward in time, reducing visual clutter (Figure 2). A tooltip also shows the actual number of patients in each group and their percentage of the total population.

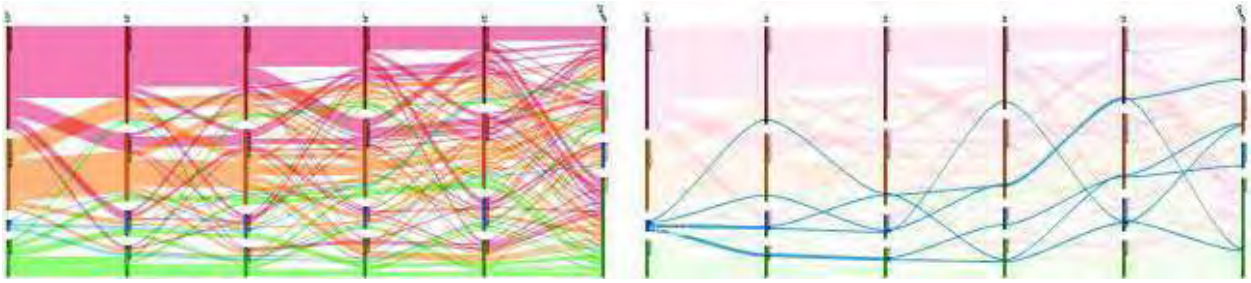


Figure 2. A 10-year range of data, sampled every two years, with forward propagation to show how the trajectories of patients change moving forward in time (left). Highlighting enables a focused view of a single category, reducing visual clutter (right).

The user can also chose between different types of coloring schemes for the paths: 1) color by the category at the first or last year (depending on the propagation direction), which shows the level of variation for a category over the length of the visualization, 2) color by transition, where the transition has a gradient from the source to target category color, which is useful for showing overall trends, and 3) color by reverse transition, where the transition path has a gradient from the target category to the source category, which is useful for category-level analysis of the distribution of source and target categories at a particular time step's category (Figure 3). To reduce visual clutter there is also an option to look at only static transitions (i.e. no change in category between time steps), and to look at only variations (i.e. only changes in the categories).

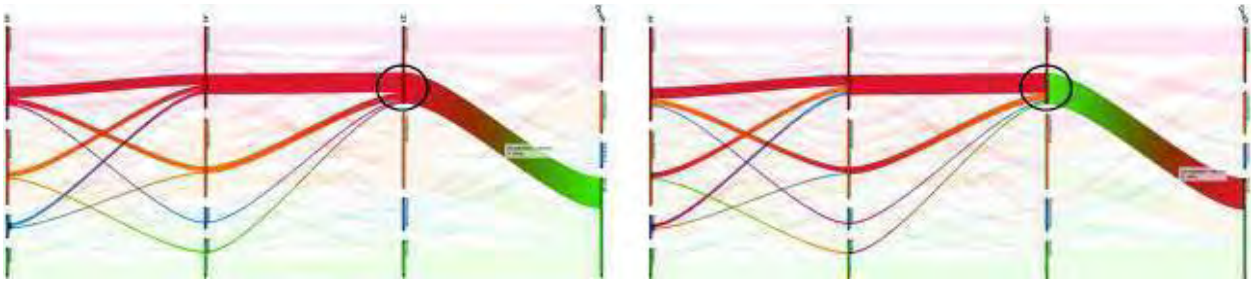


Figure 3. The user can observe separate groups by selecting individual trajectories. In addition to coloring by the starting category, paths can be colored by a gradient from source to target category (left), which redundantly encodes the category at each axis to emphasize overall trends, or by target to source category (right), which enables a rapid analysis of where paths are moving to/from at each category. The circled regions highlight this difference. On the right, it is immediately obvious what category this trajectory came from at death (Normal in green) and how this group is distributed at the previous time step.

We also include the ability to incorporate demographic data, such as gender, as additional axes (Figure 4). This feature enables the comparison of trajectories for different subpopulations based on data other than just HbA1c levels.

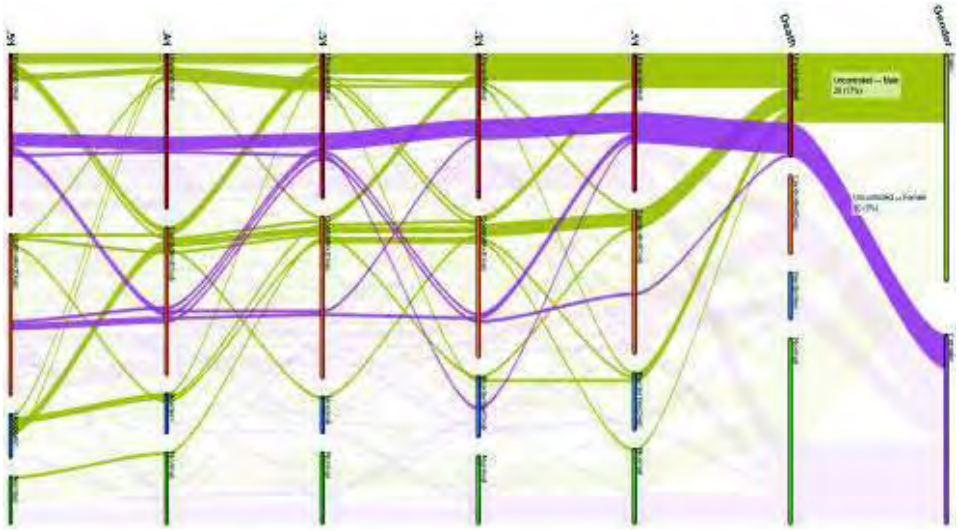


Figure 4. By adding a gender axis and selecting two groups we can compare the variability of males who were Uncontrolled at death (olive) to women who were Uncontrolled at death (purple). Men appear to have more variability over the 5-year period being visualized, as shown by the large number of transitions between different categories.

Findings

In the 10 years before death, there is a consolidating trend to improved glucose control across all diabetes control categories from uncontrolled to normal. Overall diabetes control shifts from uncontrolled diabetes for 46% of the cohort to 25% at death utilizing HbA1c and imputed glucose values. A reciprocal increase in combined borderline and normal range glucose control goes from 25% at 10 years out to 57% at death (Table 1). The trend for better glucose control is most visible in the last six months before death. The overall final glycemic trajectory is also evident in the bottom image from Figure 1 where the control category at death is colored retrospectively. Notably a

small minority of the uncontrolled sub-group remains poorly controlled over the entire disease course. By including category temporal transitions this visualization also illustrates the complexity of the underlying data, with many trajectories exhibiting a large degree of variation in HbA1c categorization over time.

Table 1. Percent of patients by Diabetes control category over 10 years prior to death using HbA1c with imputed glucose results.

Glycemic Control by HbA1c	*-10 years years to death*	*-5 years years to death*	*At Death*
Uncontrolled Diabetes	46 %	39 %	25 %
Controlled Diabetes	32 %	39 %	19 %
Borderline	5 %	11 %	12 %
Normal	17 %	12 %	45 %

Discussion

The progression of diabetes with accumulating end organ complications is well recognized. There is a clinical presumption that diabetes-related complications are also associated with worsening glycemic control for patients with end stage diabetes mellitus. Since most prospective cohort studies are organized by a patient's clinical presentation, treatment or demographics, they tend to be cross sectional studies of a population and include patients across a disease continuum. By creating a cohort organized by a death criterion with 10 or more years of diabetes lab data, we have sub-selected a general but presumably more ill diabetic population. Phasing HbA1c values by death allows data coherence that translates into the visualization of glycemic trajectories that would be less evident in cross sectional studies of diabetic patients. Understanding the course of diabetes control is important to discerning differences in outcomes, treatments and identifying sub-phenotype populations.

Death event as an organizing point for temporal data visualization permits a clear starting point to observe the course of medically treated diabetes. Cause of death is not defined, so further characterization of subpopulations visualized in the cohort, like the always uncontrolled diabetes subgroup, warrants further clinical investigation to see if they are representative of the cohort overall. All patients in this cohort had data for at least 10 years, as such our population is specific for patients under some manner of regular medical care, and interpretation of the data with respect to populations with less regular medical care should be limited. Using the imputed average glucose and average HbA1c values aligned on the cohort's endpoint enables capture of all glycemic values, including those potentially before even the diagnosis of diabetes is made.

We observed a trend to normalization of HbA1c in the last year of life. The reasons behind improved diabetes control near the end of life could include multiple factors, such as increased insulin half-life due to impaired renal and hepatic metabolism, decreased dietary intake related to anorexia or nausea, and falsely low HbA1c secondary to uremia or anemia.[19] Interestingly, the goals for end-of-life treatment in diabetic patients are generally to limit side effects of either hyper or hypoglycemia and often entail a scaling back of treatment which would be expected to be associated with more hyperglycemia not less. By using visualization tools to see the progression of HbA1c values in diabetic patients in the years before their death, our findings of glucose normalization in light of this paradigm highlight the need for further clinical investigation and interpretation.

Our data visualization tool displays temporal patterns of diabetes metric across a population and for the last years of this disease continuum. Tools such as these can only display patterns that can potentially illuminate findings that need further clinical validation and statistical investigation to determine clinical significance if any.

Future work

The visualizations we have shown here represent a small number of patients in the dataset. This has enabled us to test and refine the visualization before using large amounts of data. Next we will include diabetes-related co-morbidities, e.g. cardiovascular, neurological, and renal manifestations of prolonged diabetes illness, and additional demographic variables, e.g. age and ethnicity. We plan to link this temporal visualization to other multivariate visualizations highlighting selected groups of patients, helping to show factors related to diabetes. We are also working toward a better statistical analysis of the data, and its representation in this tool. In particular, we wish to incorporate information regarding the amount of imputed and extrapolated data in the visualization.

Conclusion

Exploring the natural disease course of diabetes control with data visualization tools permits identification of potentially clinically important trends that would be difficult to recognize otherwise. Further investigation and definition on the clinical significance of the normalization of HbA1c in the final years of life are warranted.

Acknowledgments
This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation. We acknowledge the assistance from Meghana Ganapathiraju in helping to refine the visualization. Our implementation was adapted from the d3.parsets reusable chart by Jason Davies. We also acknowledge the assistance of Mark Massing MD PhD MPH and Susan Spratt MD for comments and discussions on graphical representations of diabetes control.

References

1. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med.1993;329(14):977-986.
2. American Diabetes Association. Standards of Medical Care in Diabetes. Diabetes Care. 2009;32(S1):S13-S61.
3. The Accord Study Group. Long-term effects of intensive glucose lowering on cardiovascular outcomes. N Engl J Med. 2011;364(9):818-828.
4. The UK Prospective Diabetes Study Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet. 1998;352(9131):837-853.
5. Bendix F, Kosara R, Hauser H. Parallel sets: visual analysis of categorical data. IEEE Symp on Info Vis. 2006;12(4):133-140.
6. Riehmann P, Hanfler M, Froehlich B. Interactive Sankey diagrams. IEEE Symp on Info Vis. 2005;233-240.
7. Holman R, Paul S, Bethel M, Matthews D, Neil H. 10-year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med. 2008;359(15):1577-1589.
8. Ford E, Li C, Little R, Mokdad A. Trends in A1C concentrations among U.S. adults with diagnosed diabetes from 1999 to 2004. Diabetes Care. 2008;31(1):102-104.
9. Gebregziabher M, Egede LE, Lynch CP, Echols C, Zhao Y. Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. Am J Epidemiol. 2010;171(10):1090-1098.
10. Inselberg A, Dimsdale B. Parallel coordinates. Human-Machine Interactive Systems. 1991;199-233.
11. Hoffman H. Exploring categorical data: Interactive mosaic plots. Metrika. 2000;51(1):11-26.
12. Monroe M, Wongsuphasawat K, Plaisant C, Shneiderman B, Millstein J, Gold S. Exploring point and interval event patterns: Display methods and interactive visual query. HCIL Tech Report, University of Maryland. 2012.
13. von Landesberger T, Bremm S, Andrienko N, Andrienko G, Tekusova M. Visual analytics methods for categoric spatio-temporal data. IEEE Conf on Vis Anal Sci and Tech(VAST) 2012;183(192):14-19.
14. Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE TransVis Comput Graph, 2012;18(12):2659-2668.
15. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE guided query tool: Providing simplified access to clinical data for research and quality improvement. J Biomed Info. 2011;2:266-276.
16. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C Assay into estimated average glucose values. Diabetes Care. 2008;31(8):1473-1478.
17. Bostock M, Ogievetsky V, Heer J. D^3: Data-driven documents. IEEE Trans Visualization & Comp Graphics. 2011;17(2):2301-2309.
18. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualization. Proc. 1996 IEEE Symp Vis Lang. 1996;336-343.
19. Kalantar-Zadeh K, Derose SF, Nicholas S, Benner D, Sharma K, Kovesdy CP. Burnt-out diabetes: Impact of chronic kidney disease progression on the natural course of diabetes mellitus. J Renal Nutrition. 2009:19(1):33-37.

Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates

David Borland[1], Vivian L. West[2], W. Ed Hammond[2]
[1]RENCI, The University of North Carolina at Chapel Hill, Chapel Hill, NC;
[2]Duke Center for Health Informatics, Duke University, Durham, NC

Abstract

We present radial coordinates, a multivariate visualization technique based on parallel coordinates. The visualization contains a number of features driven by the needs of health-related data analysis, such as integrating categorical and numeric data, and comparing user-selected subpopulations via ribbon rendering. We illustrate the utility of radial coordinates by exploring primary care trust (PCT) and practice-level data from the United Kingdom's National Health Service, using three examples: lung cancer rates among PCTs, various cancer rates among only London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Introduction

With the ever-increasing size and number of health-related datasets, new analytical tools are becoming necessary to enable enhanced understanding of the vast amount of information contained within. Visualization leverages the power of the human visual system to reveal patterns and relationships in data by mapping the data to visually salient features.

One of the challenges for visualization of health-related data is the desire to incorporate data of many types (e.g. lab results, demographics, medications, vital signs, and genomic data) from various sources. We have developed a multivariate visualization technique, radial coordinates, that enables visual analysis of a wide range of health-related datasets and handles both numeric and categorical data (Figure 1). Radial coordinates facilitates the interactive exploration of datasets to reveal patterns in the data, discover relationships between variables, and compare user-defined subpopulations. In this manner we support the pursuit of hypothesis formations that can elicit further inquiry and lead to new knowledge.

An overview of an initial radial coordinates prototype applied to query data was given previously.[1] In this paper we provide a more in-depth description of the various features of a new implementation, which includes several new features, and discuss its application to primary care trust (PCT) and practice-level data from the National Health Service (NHS) in the United Kingdom (UK). We present three examples illustrating the use of radial coordinates to explore the NHS data: lung cancer rates among PCTs, a comparison of various cancer rates among London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Previous Work

Our visualization is based largely on parallel coordinates, a multivariate visualization technique which represents each dimension as a parallel axis, and each data entity as a line connecting the entity's value at each axis.[2,3] Non-parallel arrangements of axes have also been investigated.[4] Our radial coordinates arrangement differs in that the radial layout maintains a square aspect ratio even with many axes, and enables utilization of the space in the center of the radial layout. Parallel coordinates have been combined with various other visualization techniques[5-7], including direct integration of scatter plots.[8,9] In our visualization we include a scatter plot based on the first two principal components to enhance the ability to find clusters in high-dimensional data in an intuitive manner (Figure 1a). Future work will explore combinations with other techniques. We also incorporate chords representing the correlations between axes in a manner similar to Circos.[10] Extensions to parallel coordinates for incorporating categorical data include parallel sets[11] and hammock plots.[12] Both represent multiple data points as paths between axes, with the number of data points encoded as path width. Our curve spreading technique incorporates categorical and continuous data while still enabling the visualization of individual data points (Figure 2). Various techniques have been developed to combine multiple data points to enhance the understanding of large datasets[13,14] and observe clusters via edge bundling techniques.[15,16] Our ribbon rendering technique enables enhanced visualization of user-selected data points, including overlaying information of statistical data (median value and quartile ranges) of interest to the health-care community (Figure 1b). Axis ordering is an important element of parallel coordinates visualizations, as it is typically easier to notice relationships between variables with adjacent axes.[17-19] We employ a correlation-based clustering technique and also introduce dynamic reordering of categorical axis values to cluster similar values based on user-defined selections (Figures 3c, 3d).

Methods

Data

PCTs, abolished in 2013 due to NHS reorganization, were regional administrative bodies in the UK responsible for commissioning health services from providers and providing community health services. Here we investigate 26 variables measuring various health and socioeconomic factors for 147 of the 152 PCTs in England (five were removed due to missing data). Health factors include cancer rates, drug prescription rates, and factors related to diabetes prevalence and treatment. Socioeconomic factors include socioeconomic deprivation, economic output, geographic region, and local region classification (e.g. *Manufacturing Towns* and *Coastal and Countryside*) from the Office for National Statistics (ONS).

We also demonstrate our visualization with data showing the prevalence of a number of medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). There were ten SHAs in England from 2006-2013.

Visualization

The radial coordinates visualization, implemented using the D3 JavaScript library[20], represents each variable in a multivariate dataset by an axis, with the axes arranged radially around a circle. Each individual entity is represented by a curve that connects the value of that entity at each axis. Figure 1 gives an example applied to PCT data, with four PCT curves highlighted in red by the user.

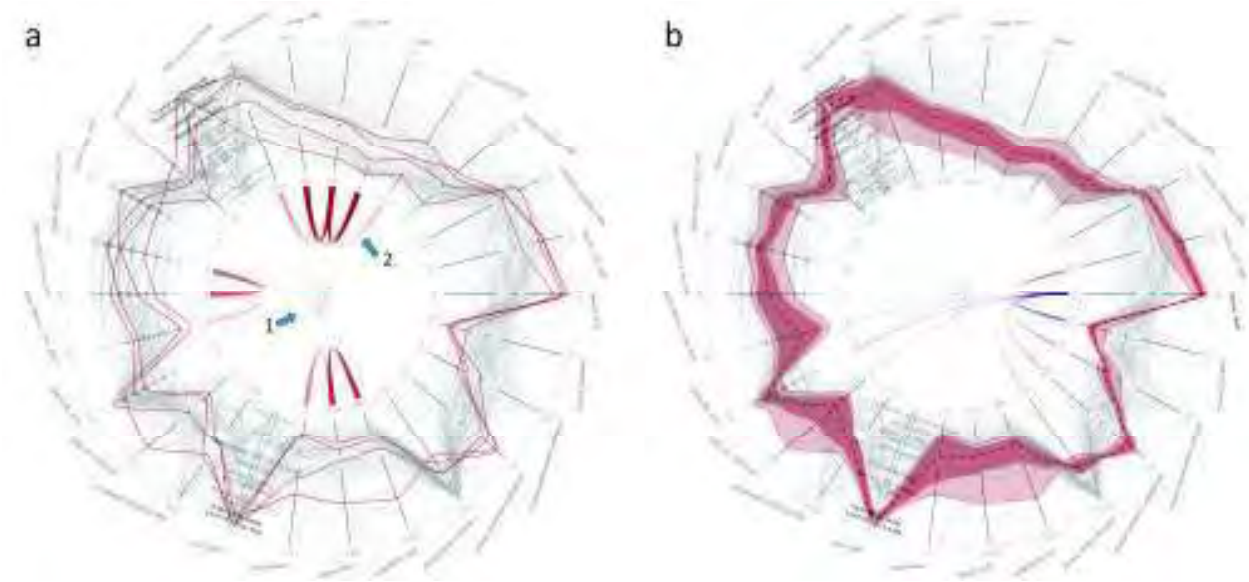


Figure 1. Radial coordinates visualizations of NHS PCT data. User-highlighted curves (red) enable the comparison of four PCTs across multiple variables (a). A linked scatterplot of the first two principal components can help show clusters in high-dimensions (a1). Chords connecting axes represent correlations (positive: red, negative: blue) above a user-defined threshold (a2). Ribbon rendering enables a simplified representation of user-defined subpopulations, displaying the data range optionally overlaid with median value and inner quartile ranges (b). Mouse over of an axis shows all correlations with that axis, regardless of user-defined threshold (b).

User selection of individual curves enables a visual comparison of how different entities relate across the various axes. A radial layout elegantly handles large numbers of axes while maintaining a square aspect ratio, also enabling the use of the center of the layout for supplemental visualizations, such as axis correlation chords and a scatterplot of the first two principal components (Figure 1a). Ribbon rendering uses a sliding window algorithm to draw the area between the innermost and outermost boundary of selected curves in a semi-transparent solid color, making it easier to see the spread of each subpopulation. An optional summary statistic overlay shows the inner quartile range and median value of each subpopulation (Figure 1b). Other visualization features include data-type dependent axis distribution visualizations and curve spreading for categorical and discrete data (Figure 2).

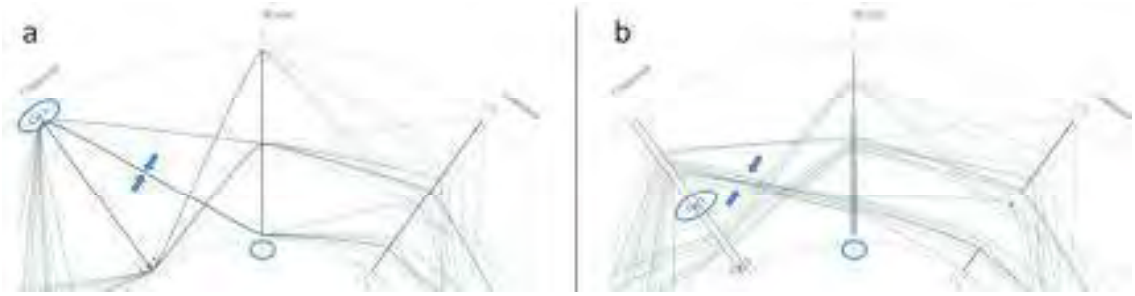


Figure 2. A sample data set without (a) and with (b) data-type dependent axis distribution visualizations and curve spreading. Axis distribution visualizations represent categorical axes as a stacked bar chart, discrete numeric axes as a histogram, and continuous numeric axes as a quartile plot[21], enabling rapid evaluation of the data type and overall distribution of the data for each axis. Curve spreading for categorical and discrete axes enables improved visualization of individual curves and clusters of curves, such as the number of data points with a Categorical value of Cat 1 and a Discrete value of three (highlighted in blue).

Results

Lung Cancer Prevalence

In Figure 3 the user has clicked on the lung cancer rate axis (*lung_Combined_DSR*), causing PCTs in the upper quartile of lung cancer rate to be automatically colored red, and the lower quartile blue. High and low lung cancer rates can now be compared across all dimensions in the data (Figure 3a). In the upper portion of the visualization it is apparent that PCTs with high and low lung cancer rates also tend to have high and low values for *extent*, *average_score*, *average_rank*, and *local_concentration* (also indicated by the correlation chords connecting these axes), which represent measures of social deprivation (poverty rate, socioeconomic status, etc.)

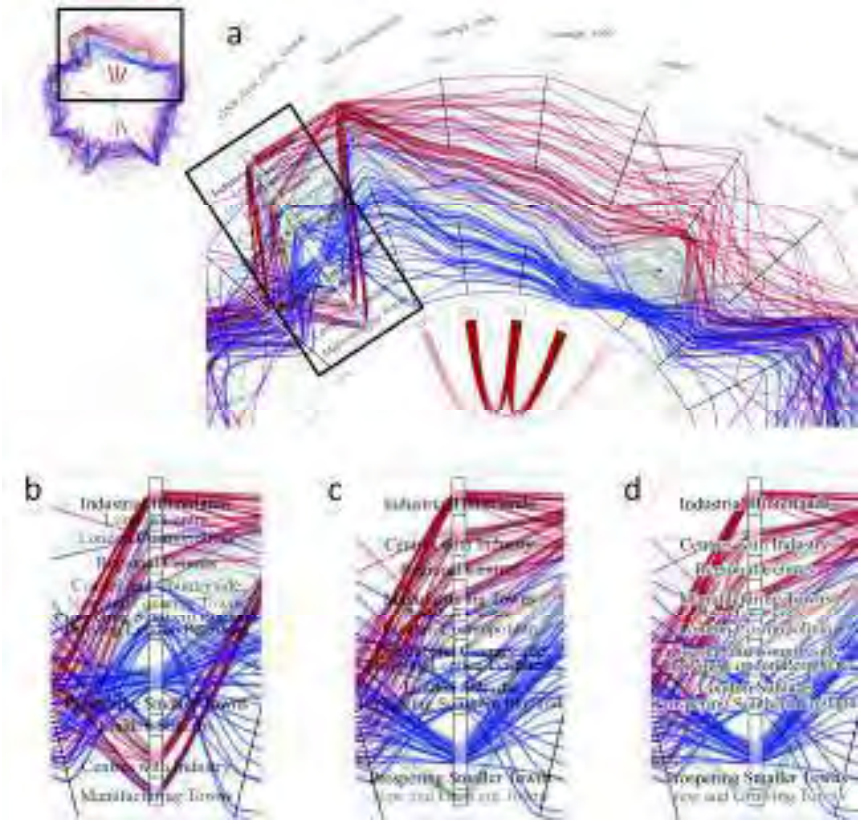


Figure 3. Visualization of lung cancer rates (red = upper quartile, blue = lower quartile) in 147 primary care trusts (PCTs) in the UK. High and low lung cancer rates tend to cluster based on regional classification (b), made clearer with automatic categorical axis reordering to cluster similar regions (c, d).

The red and blue curves also form clusters on the *ONS_Area_Class_Group* axis, a local region categorization from the ONS. Investigating this axis (Figures 3b-d) shows that *Industrial Hinterlands*, *Centres with Industry*, *Regional Centers*, and *Manufacturing Towns* all have high lung cancer rates, whereas *Prospering Smaller Towns*, *Prospering Southern England*, *London Suburbs*, and *Thriving London Periphery* all have low lung cancer rates. The discovery of such relationships via exploring the data visually drives the formation of causal hypotheses (e.g. pollution levels or smoking prevalence), which can be investigated further.

London Suburb Comparison

In Figure 4a a single PCT, Harrow, was seen to have the lowest lung, bladder, and colorectal cancer rates compared to all other PCTs, and has been highlighted in red. Harrow is classified as a *London Suburb*, so in Figure 4b the user has highlighted the other London suburbs in blue for comparison, made easier in Figure 4c via ribbon rendering. Harrow is shown to have a much higher value for the *oesophageal_Combined_SRR* axis, and thus a much higher esophageal cancer rate, than the other London suburbs, which are almost all in the lower quartile. This visualization raises the question of why Harrow has such a disparity in the rates of different cancers.

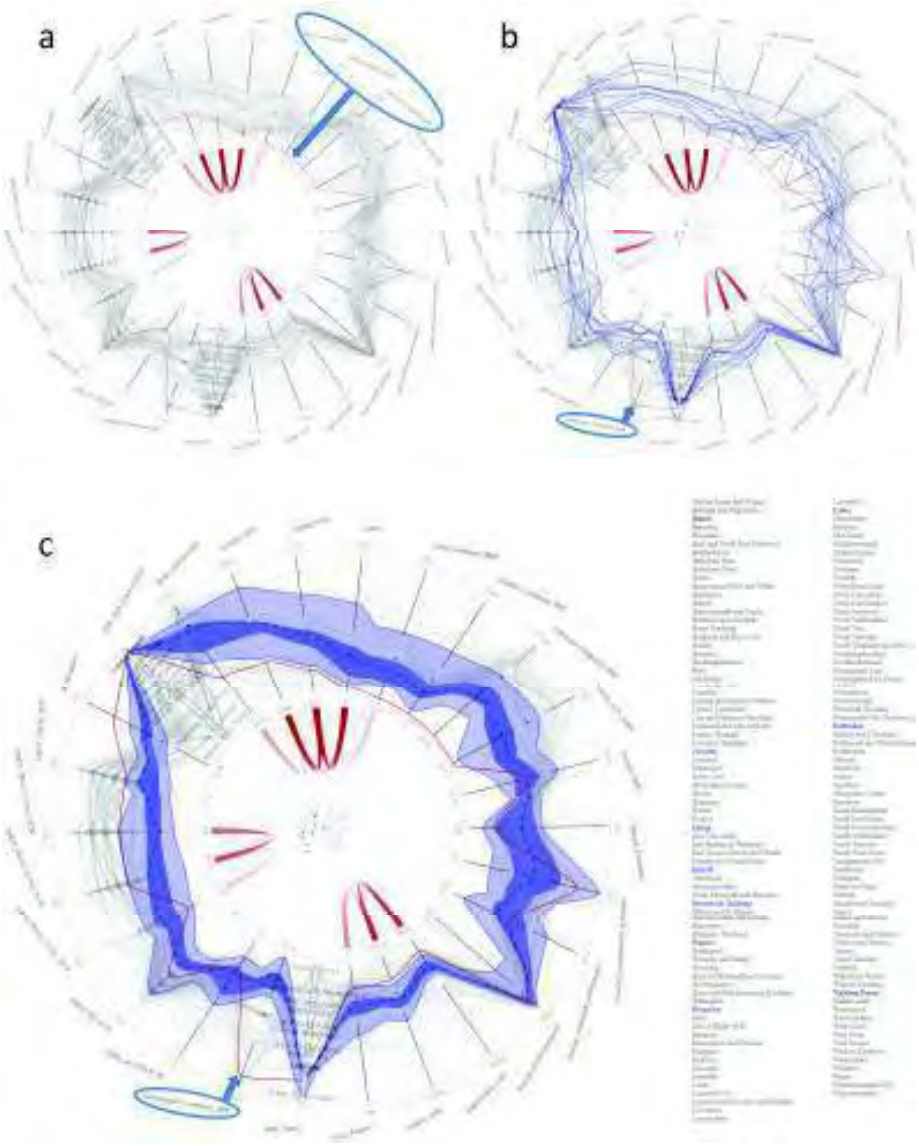


Figure 4. The Harrow PCT (red) has the lowest lung, bladder, and colorectal cancer rates (circled) among all 147 PCTs in the NHS dataset (a). Comparing Harrow to the other London suburbs (blue) reveals that Harrow has a much higher esophageal cancer rate (circled) than the other suburbs (b). Ribbon rendering makes it easier to visually compare Harrow with the other London suburbs (c).

According to the 2011 Census[22] Harrow is very diverse, with 63.8% of its population from the Black and Minority Ethnic communities, including the highest concentration of Sri Lankan Tamils and Gujarati Hindus in the UK and Ireland. India is known to have relatively low cancer rates in general, but some of the highest rates for oral and esophageal cancers in the world[23], which may help explain this phenomenon. Although further analysis is necessary, this example shows the utility of radial coordinates and ribbon rendering to compare subpopulations.

Practice-Level Data

Figures 8a and 8b show the prevalence of various medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). Figure 8a highlights in red two practices that appear to be outliers in the PCA scatterplot. Ribbon rendering makes apparent that they have the two highest prevalences for 12 of the 21 medical problems represented in the data. Figure 8b applies ribbon rendering to the remaining 1502 practices, making it easier to compare maximum and minimum values of medial problem rates for the two subpopulations.

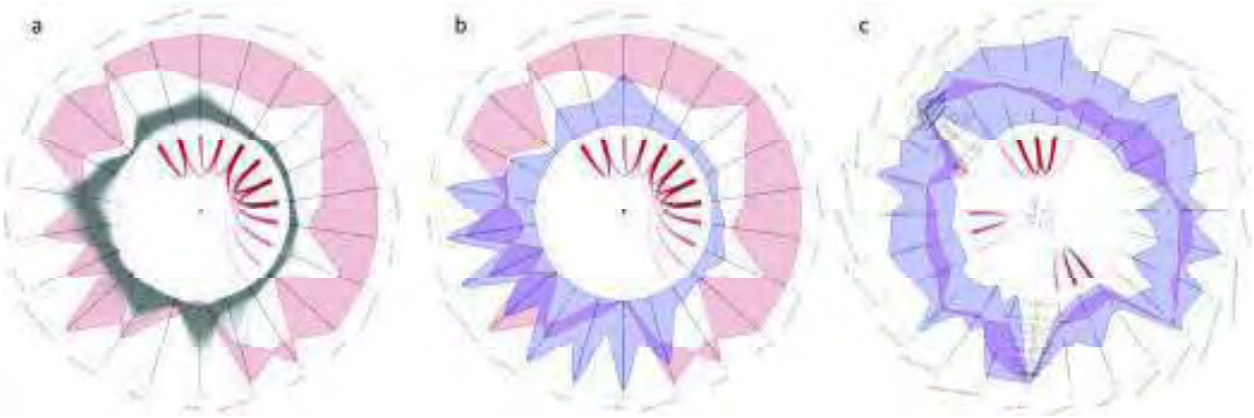


Figure 8. Two out of the 1504 practices in the London SHA, highlighted in red, have the two highest prevalences for 12 of the 21 medical problems represented in the NHS practice-level data (a and b). Comparing the PCTs containing these practices (red) to all other London PCTs (blue) does not reveal any major differences (c).

The two practices highlighted in red are Royal Hospital Chelsea in the Kensington and Chelsea PCT, and Nightingale House in the Wandsworth PCT. Because these two practices stood out so dramatically in the practice-level data, the user performed a PCT-level comparison of all London PCTs (Figure 8c). Interestingly, the Kensington and Chelsea and the Nightingale House PCTs (red) do not appear very different when compared to the other London PCTs (blue). Further research determined that Royal Hospital Chelsea is a retirement and nursing home for British soldiers and Nightingale House is a nursing home for the Jewish community that specializes in dementia, which may explain the high prevalence of problems such as dementia, hypertension, stroke, heart failure, and cancer in these two practices.

Conclusion

We have presented radial coordinates, a multivariate visualization technique based on parallel coordinates that incorporates features, such as per-axis population distribution visualizations based on data type (continuous, discrete, and categorical), direct visualization of correlations between variables, curve spreading for discrete and categorical data, visualization of summary statistics for user-selected subpopulations via ribbon rendering, and automatic reordering of categorical values based on user selection, driven by the needs of health-related data visualization.

We have applied radial coordinates to data from the UK's NHS at both the PCT and individual practice levels. Visualization of lung cancer rates among PCTs discovered possible relationships among lung cancer rate, socioeconomic factors, and regional classification. A comparison of London suburb PCTs revealed a potentially interesting PCT with a much higher esophageal cancer rate than other similar PCTs. Visualizing medical problem prevalence among over 1500 London practices showed two practices that have much higher rates of many medical problems. These examples illustrate the utility of the combination of visualization techniques embodied in our radial coordinates tool, and underline the need for further research in the use of visualization to aid in the analysis of complicated health-related datasets.

Acknowledgments

NHS and other UK data were made available courtesy of the BT Health Cloud. This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

References

1. West V, Borland D, Hammond WE. Visualization of EHR and Health Related Data for Information Discovery. In Proceedings of the 2013 AMIA Workshop on Visual Analytics in Healthcare. November 2013.
2. Gannet H. General summary, showing the rank of states, by ratios. 1880.
3. Inselberg A. The plane with parallel coordinates. Visual Computer. 1985;1(4):69-91.
4. Tominiski C, Schumann H. An event-based approach to visualization. In Proceedings of the Eighth International Conference on Information Visualization (IV'04). July 2004;101-107.
5. Rodrigues Jr. JF, Traina AJM, Traina Jr. C. Frequency plot and relevance plot to enhance visual data exploration. In Proceedings of the XVI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'03). 2003;117-124.
6. Edsall RM. The parallel coordinate plot in action: Design and use for geographic visualization. Computational Statistics and Data Analysis. 2003;43(4):605-619.
7. Siirtola H. Combining parallel coordinates with the reorderable matrix. In Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization. July 2003;63-74.
8. Holten D, van Wijk JJ. Evaluation of cluster identification performance for different PCP variants. Computer Graphics Forum. 2010;29(3):793-802.
9. Harter JM, Wu X, Alabi OS, Phadke M, Pinto L, Dougherty D, Petersen H, Bass S, Taylor II RM. Increasing the perceptual salience of relationships in parallel coordinate plots. In Proceedings of SPIE Visualization and Data Analysis 2012. January 2012.
10. Krzywinksi M, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Research. September 2009;19(9):1639-1645.
11. Kosara R, Bendix F, Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics. July/August 2006;12(4):558-568.
12. Schonlau M. Visualizing categorical data arising in the health sciences using hammock plots. In Proceedings of the Section on Statistical Graphics, American Statistical Association. 2003.
13. Fua YH, Ward MRE. Hierarchical parallel coordinates for exploration of large datasets. In Proceedings of the Conference on Visualization '99: Celebrating Ten Years. 1999;43-50.
14. Heinrich J, Weiskopf D. Continuous parallel coordinates. IEEE Transactions on Visualization and Computer Graphics. 2009;15(6):1531-1538.
15. Zhou H, Yuan X, Qu H, Cui W, Chen B. Visual clustering in parallel coordinates. Computer Graphics Forum. May 2008;27(3):1047-1054.
16. Heinrich J, Luo Y, Kirkpatrick AE, Zhange H, Weiskopf D. Evaluation of a bundling technique for parallel coordinates. In Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications. 2012;594–602.
17. Ankerst M, Berchtold S, Keim DA. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In Proceedings of the IEEE Symposium on Information Visualization. 1998;52-60.
18. Peng W, Ward MO, Rundensteiner EA. Clutter reduction in multi-dimensional data visualization using dimension reordering. In Proceedings of the IEEE Symposium on Information Visualization. 2004;89-96.
19. Seo J, Shneiderman B. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In Proceedings of the IEEE Symposium on Information Visualization. 2004;65-72.
20. Bostock M, Ogievetsky V, Heer J. D3: Data-driven documents. IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis). 2011;17(12).
21. Tufte ER. The Visual Display of Quantitative Information. 2nd ed. Cheshire, CN:Graphics Press. 2001.
22. Office for National Statistics. 2011 Census: Ethnic group, local authorities in England and Wales. 2012.
23. Sinha R, Anderson DE, McDonals SS, Greenwald P. Cancer risk and diet in India. Journal of Postgraduate Medicine. July-September 2003;49(3).

Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels

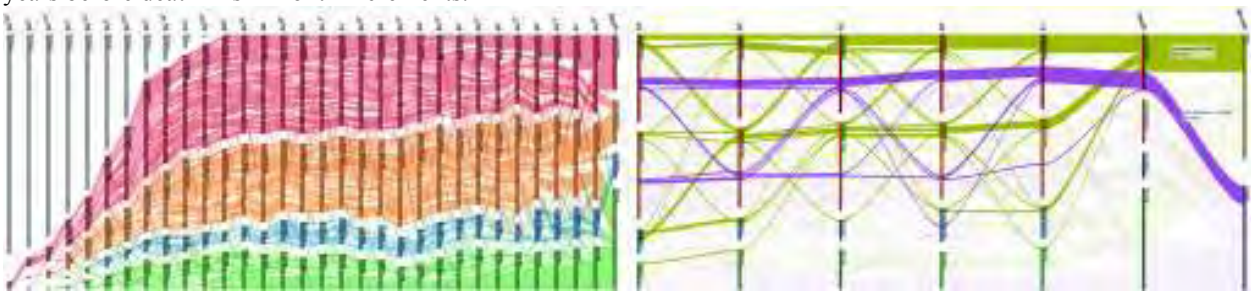
Hina Shah[1], David Borland[2], Eugenia McPeek Hinz[3], Vivian L. West[1], W. Ed Hammond[1]
[1]Duke Center for Health Informatics, Duke University, Durham, NC;
[2]RENCI, The University of North Carolina at Chapel Hill, Chapel Hill, NC;
[3]DHTS Duke Medicine, Durham, NC

Introduction

In this demonstration we present a visualization tool for a cohort of patients with diabetes (via ICD9 codes) from Duke University's data warehouse, visualizing their Hemoglobin A1c (HbA1c) levels over time, aligned by death, to explore trajectories of glycemic control. To the best of our knowledge, temporal visualization of glycemic control for a diabetic population standardized on death has not previously been presented. Our visualization groups HbA1c values into ordered categories of glycemic control, utilizing a method based on parallel sets and Sankey diagrams to view temporal patterns in HbA1c values. We incorporate a number of features for interactive data exploration like: viewing the progression of values either forwards or backwards in time, highlighting multiple subpopulations, coloring based on the category along each path in the data or at the beginning/end of each path, and the incorporation of demographic data, such as gender.

Methods

Data from Duke University's data warehouse were extracted using DEDUCE, an electronic health record (EHR) query tool developed at Duke University. The final cohort includes data from 121 patients with diabetes mellitus (with and without complications), a death indicator, prescribed antihyperglycemics, and at least 10 years of HbA1c laboratory values. We average HbA1c values over 6 month time intervals. In the case of missing HbA1c values within a 6 month period, we first attempt to impute a HbA1c value from average glucose (AG) values over that period of time if available, otherwise the previous HbA1c value (measured or imputed) is carried forward. HbA1c values are then categorized based on the severity of diabetes: Normal < 5.7, Borderline $[5.7, 6.5)$, Controlled $[6.5, 8)$, and Uncontrolled ≥ 8. The sampled data is time-aligned by the death event for each patient. The visual representation of diabetes progression propagates backwards in time initially. Time is represented as number of years before death in six month increments.



(a) (b)

Figure 1. Diabetes progression visualizations without and with a gender axis: (a) Overview visualization with paths colored by the current HbA1c at each time step, useful for emphasizing overall temporal trends. (b) Adding a gender axis and selecting two groups, we can compare the variability of males who were Uncontrolled at death (olive) to women who were Uncontrolled at death (purple). Men appear to have more variability over the 5-year period being visualized, as shown by the large number of transitions between different categories.

Our visualization tool was developed using the D3 JavaScript library. The aim of this visualization is to investigate temporal trajectories of HbA1c levels for a large cohort of diabetes patients over a number of years prior to death. Parallel sets is chosen for showing HbA1c summary trajectories. Each vertical axis is a time step. The user can choose the frequency of these time steps, with a minimum sampling frequency of six months, and also the maximum number of years before death. The death event axis is placed at the right with all other time steps moving backwards in time to the left (Figure 1). Each vertical axis is split into the four HbA1c categories (Normal in green, Borderline in blue, Controlled in orange, and Uncontrolled in red), and a Missing category in grey (for patients with more than 10 years of data). The height of each axis category represents the proportion of the patients in that category at that point in time. Paths moving between axes recursively split, moving backwards from death to show the trajectories of similar groups of patients. The visualization can show trends either starting at the death event, i.e. going backwards

in time, or starting at the last year in the visualization, i.e. going forward in time. The user can highlight one or more groups of patients by clicking on categories or trajectories to highlight the behavior of that group of patients going backward and forward in time, reducing visual clutter (Figure 2). We also include the ability to incorporate demographic data, such as gender, as additional axes (Figure 1). This feature enables the comparison of trajectories for different subpopulations based on data other than just HbA1c levels.

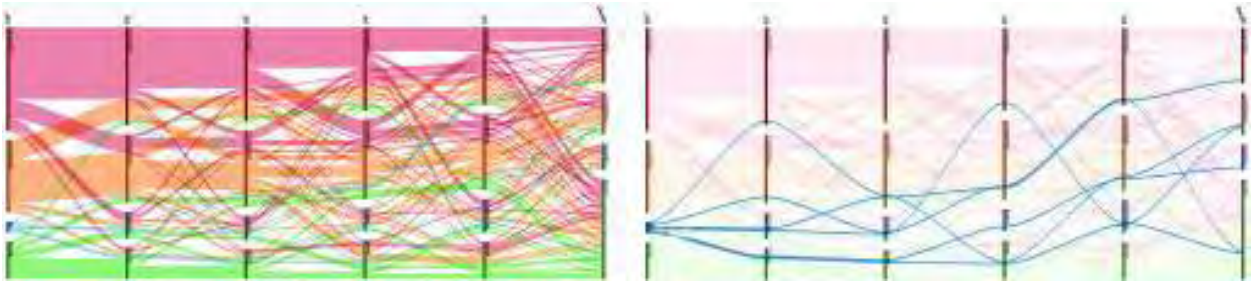


Figure 2. A 10-year range of data, sampled every two years, with forward propagation to show how the trajectories of patients change moving forward in time (left). Highlighting enables a focused view of a single category, reducing visual clutter (right).

The user can also chose between different types of coloring schemes for the paths: 1) color by the category at the first or last year (depending on the propagation direction), which shows the level of variation for a category over the length of the visualization; 2) color by transition, where the transition has a gradient from the source to target category color, useful for showing overall trends; and 3) color by reverse transition, where the transition path has a gradient from the target category to the source category, useful for category-level analysis of the distribution of source and target categories at a particular time step's category (Figure 3). To reduce visual clutter, there is also an option to look at only static transitions (i.e. no change in category between time steps), and to look at only variations (i.e. only changes in the categories).

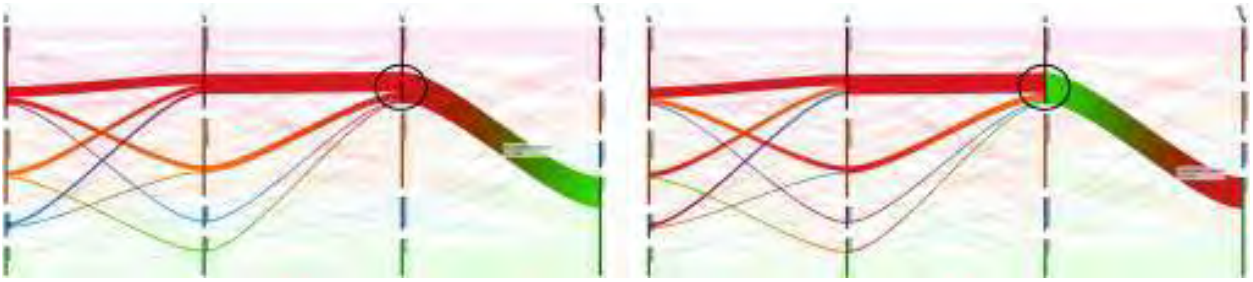


Figure 3. In addition to coloring by the starting category, paths can be colored by a gradient from source to target category (left), which redundantly encodes the category at each axis to emphasize overall trends, or by target to source category (right), which enables a rapid analysis of where paths are moving to/from at each category. The circled regions highlight this difference. On the right, it is immediately obvious what category this trajectory came from at death (Normal in green) and how this group is distributed at the previous time step.

Acknowledgments

This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation. We acknowledge the assistance from Meghana Ganapathiraju in helping to refine the visualization. Our implementation was adapted from the d3.parsets reusable chart by Jason Davies.

References:

1. Gebregziabher M, Egede LE, Lynch CP, Echols C, Zhao Y. Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. Am J Epidemiol. 2010;171(10):1090-1098
2. Bendix F, Kosara R, Hauser H. Parallel sets: visual analysis of categorical data. IEEE Symp on Info Vis. 2006;12(4):133-140.
3. Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE TransVis Comput Graph, 2012;18(12):2659-2668.

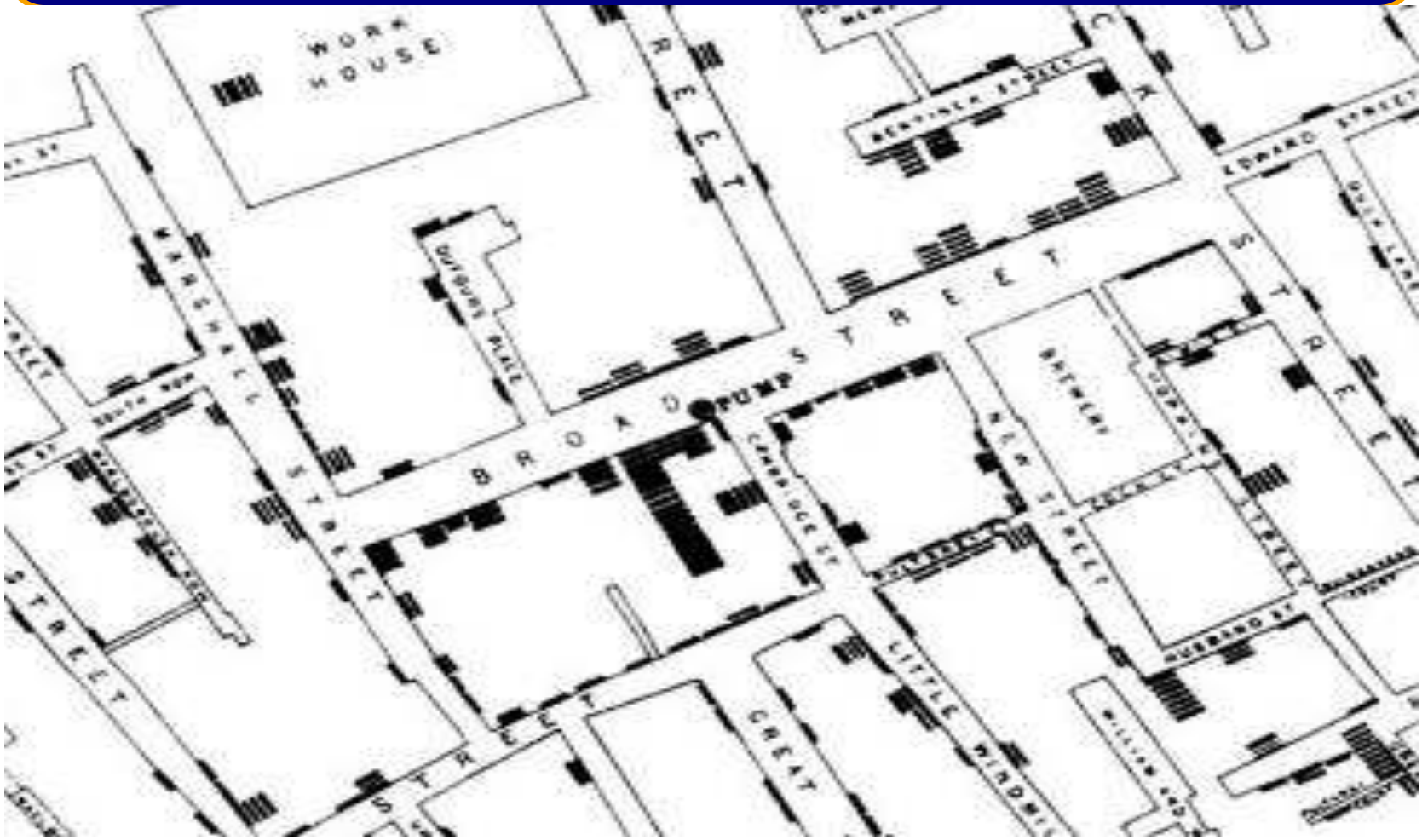
Exploring Novel Visualizations of Electronic Health Record Data

Meghana Ganapathiraju[1]
Mentors: David Borland, PhD[2], Vivian West, PhD[3], W.Ed Hammond, PhD[3]
[1]Green Hope High School, [2]RENCI, [3]Duke University

Information Visualization

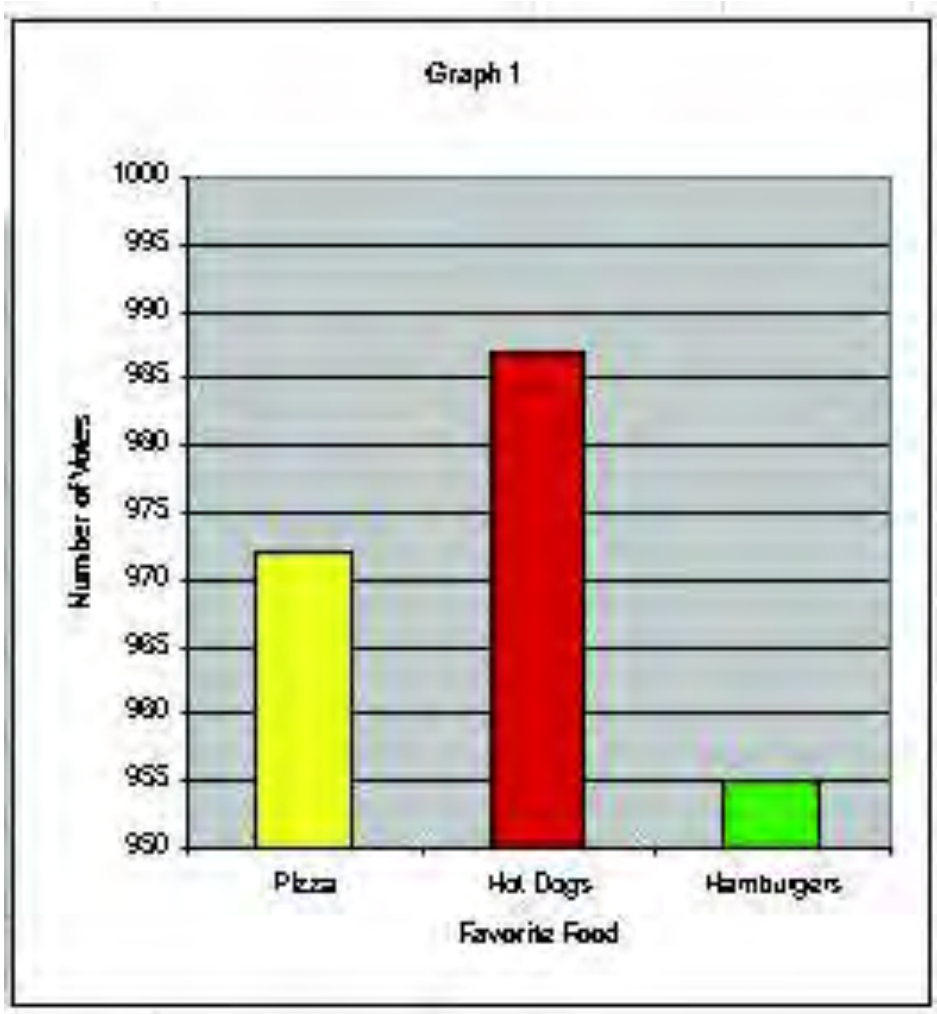
- The use of interactive visual representations of abstract data to amplify cognition (Ware, 2004)
- Makes interpretation of information easier
- Must correctly represent information or the visualization can be misleading or confusing
- Visualization can help detect key patterns otherwise difficult to pick out

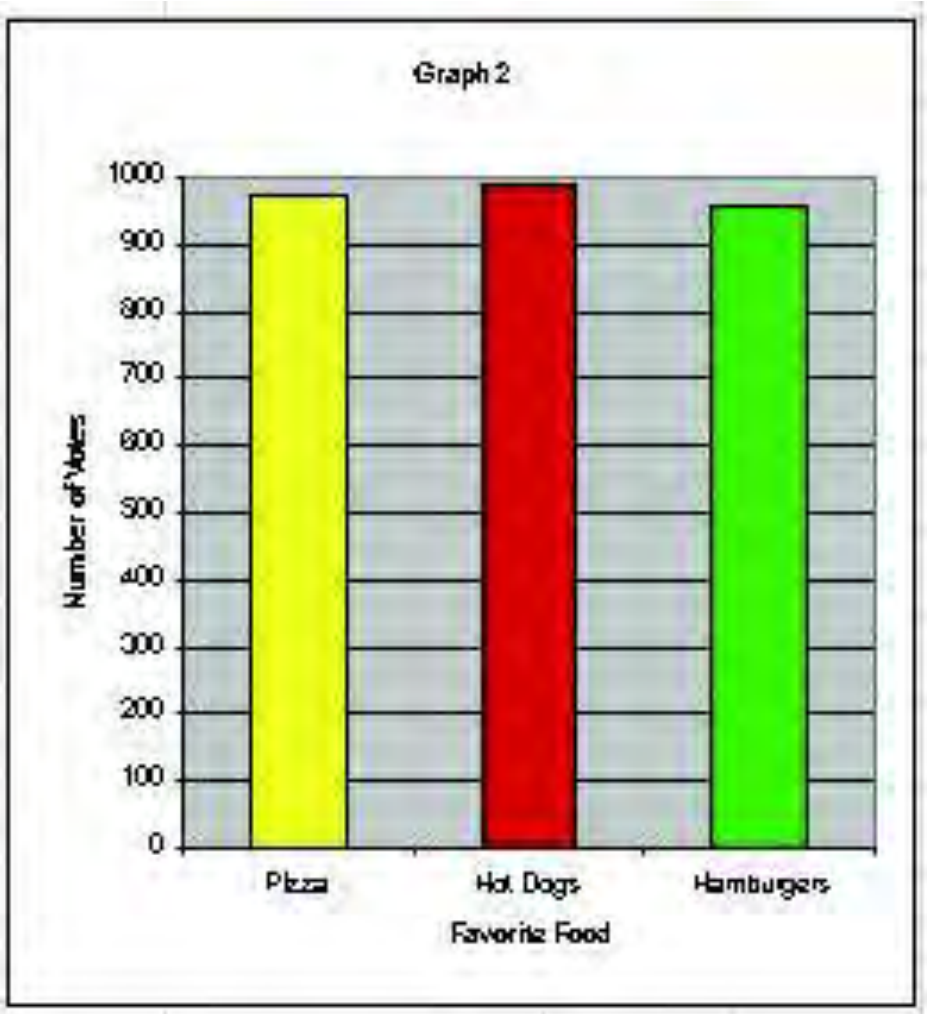
John Snow's Cholera Map



http://flowingdata.com/2007/09/12/john-snows-famous-cholera-map/

Information Visualization





Visualizing Multivariate Data

- Multivariate data has more than 2 or 3 dimensions
- Lose information if shown in traditional graphs/ plots
- New methods of visualization are necessary

Multivariate visualization techniques

- Scatterplot Matrix
- Starplots
- Parallel sets/coordinates

Purpose

- Visualize EHR data
  - Systems are increasing in size and complexity
  - Researchers se the database in different ways

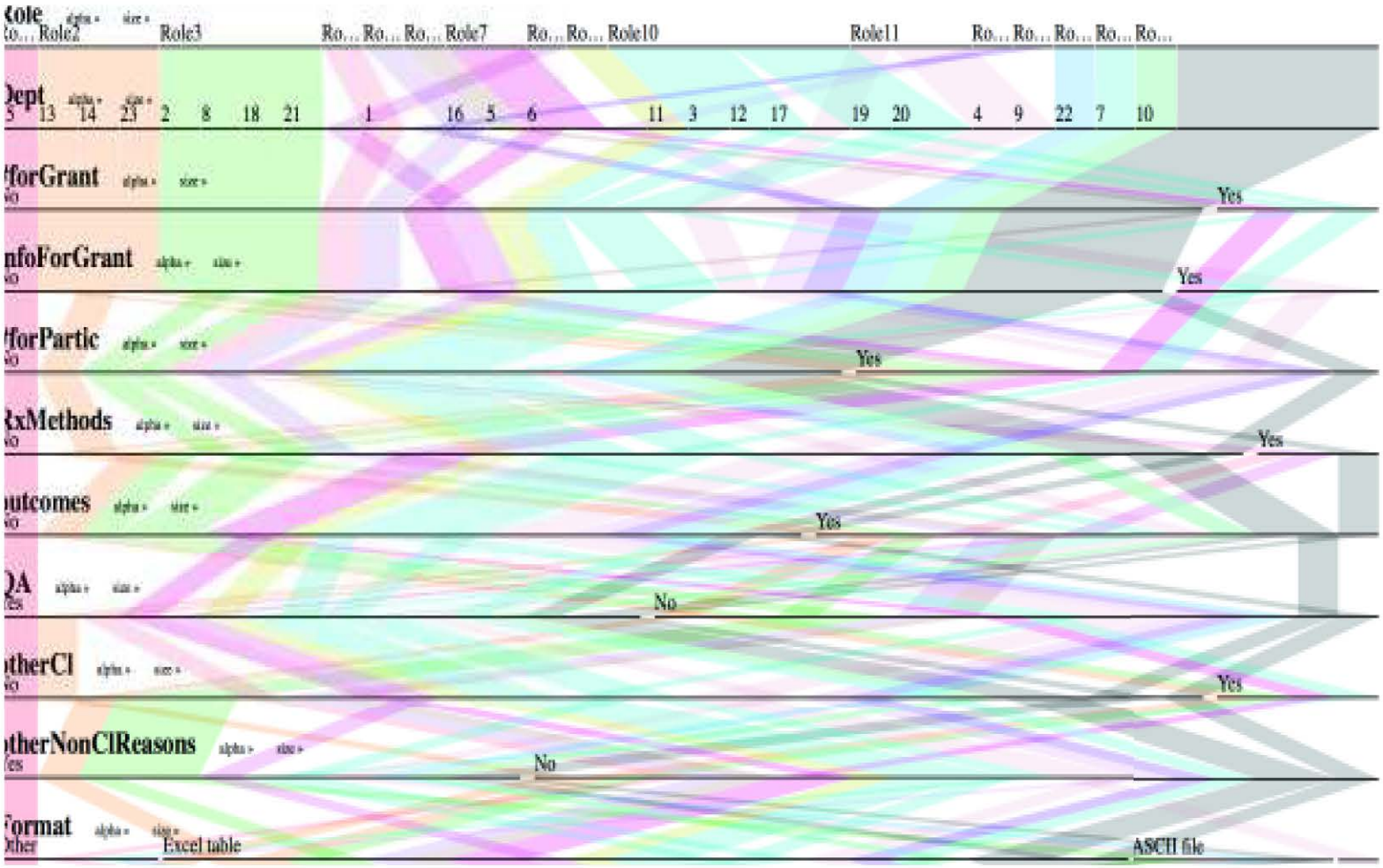
- Visual representations may facilitate additional insight

Methods

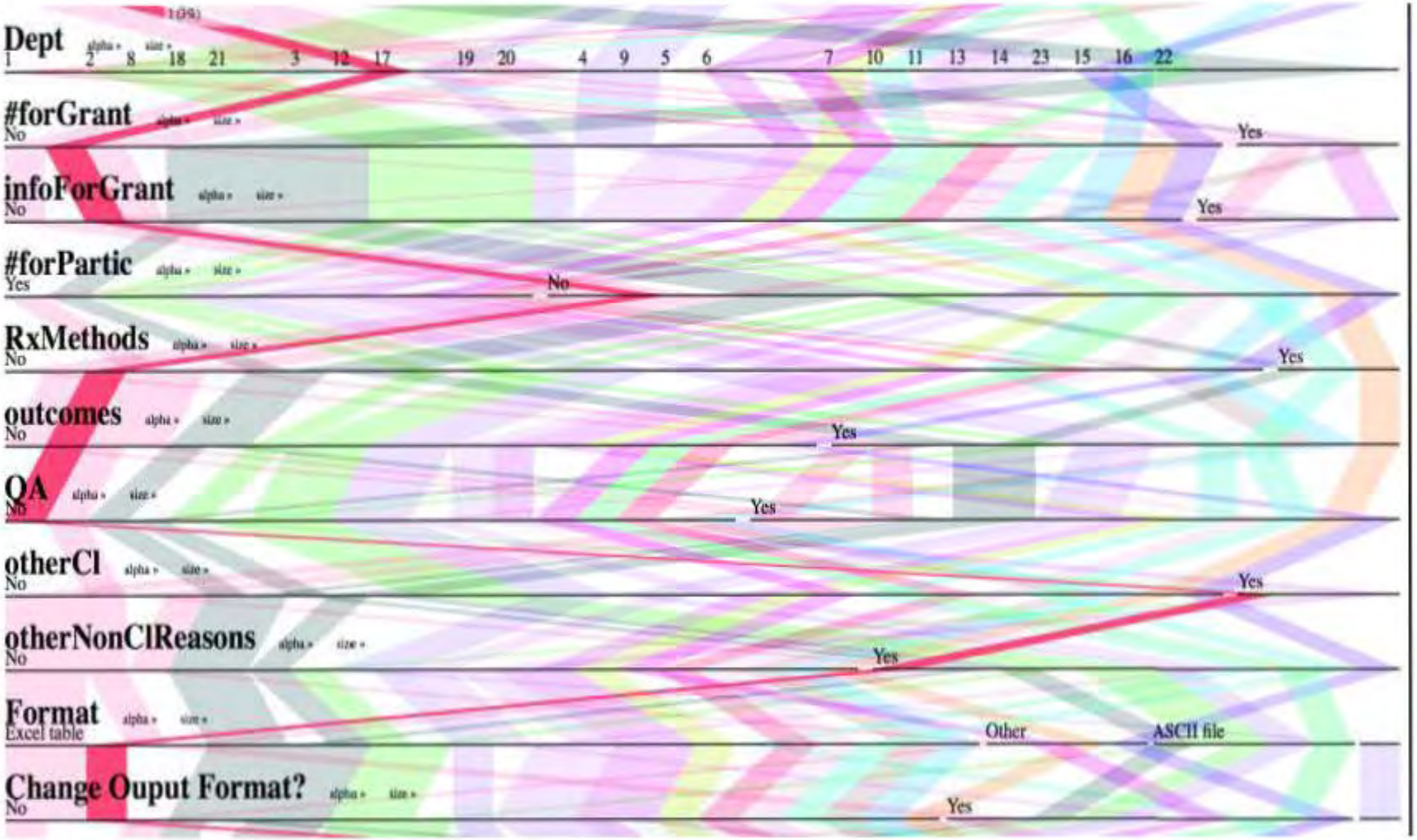
- Surveyed researchers
- Data stored as an excel file
- D3 JavaScript library used to produce the visualization

Parallel Sets Visualization

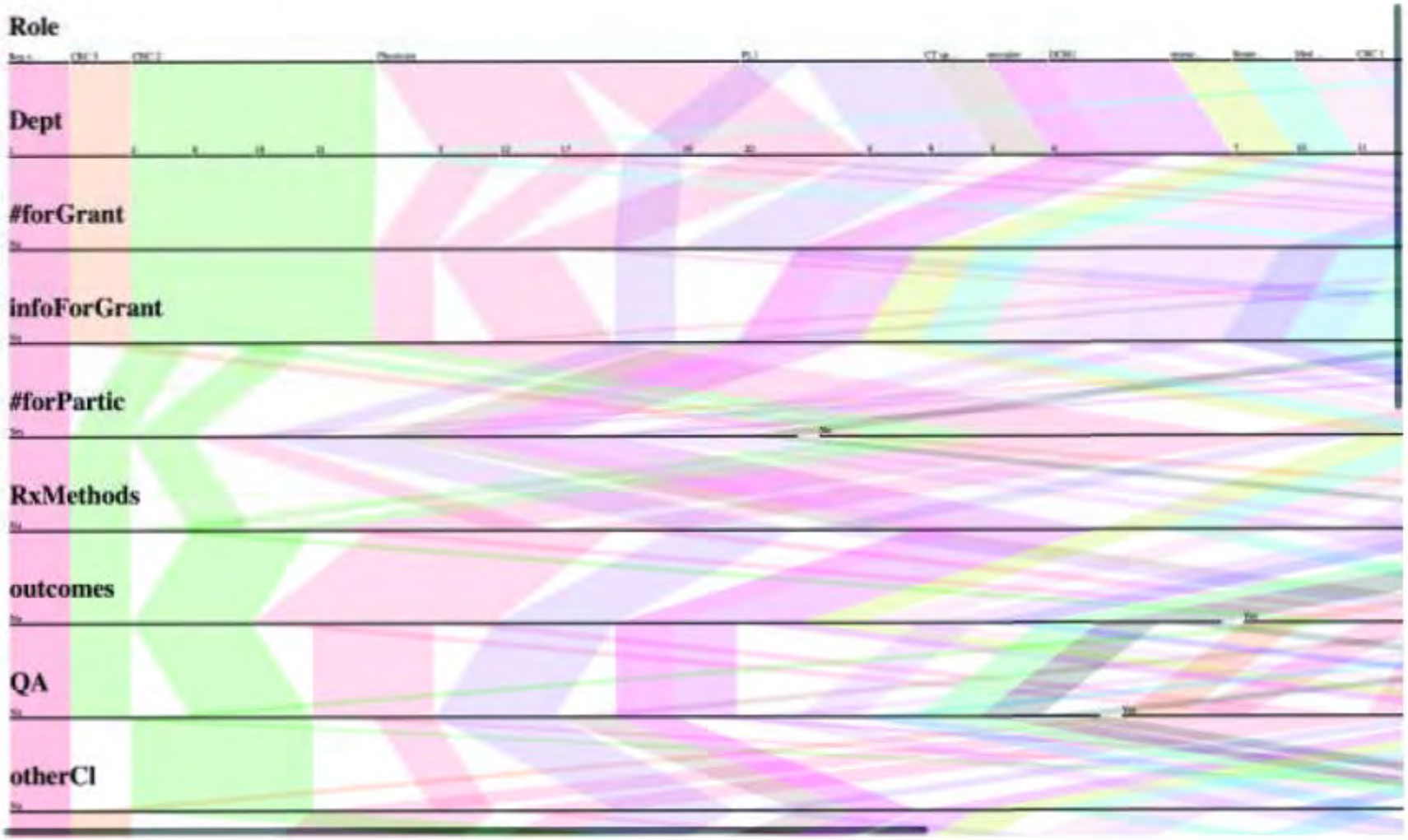
[Live Demo Link]



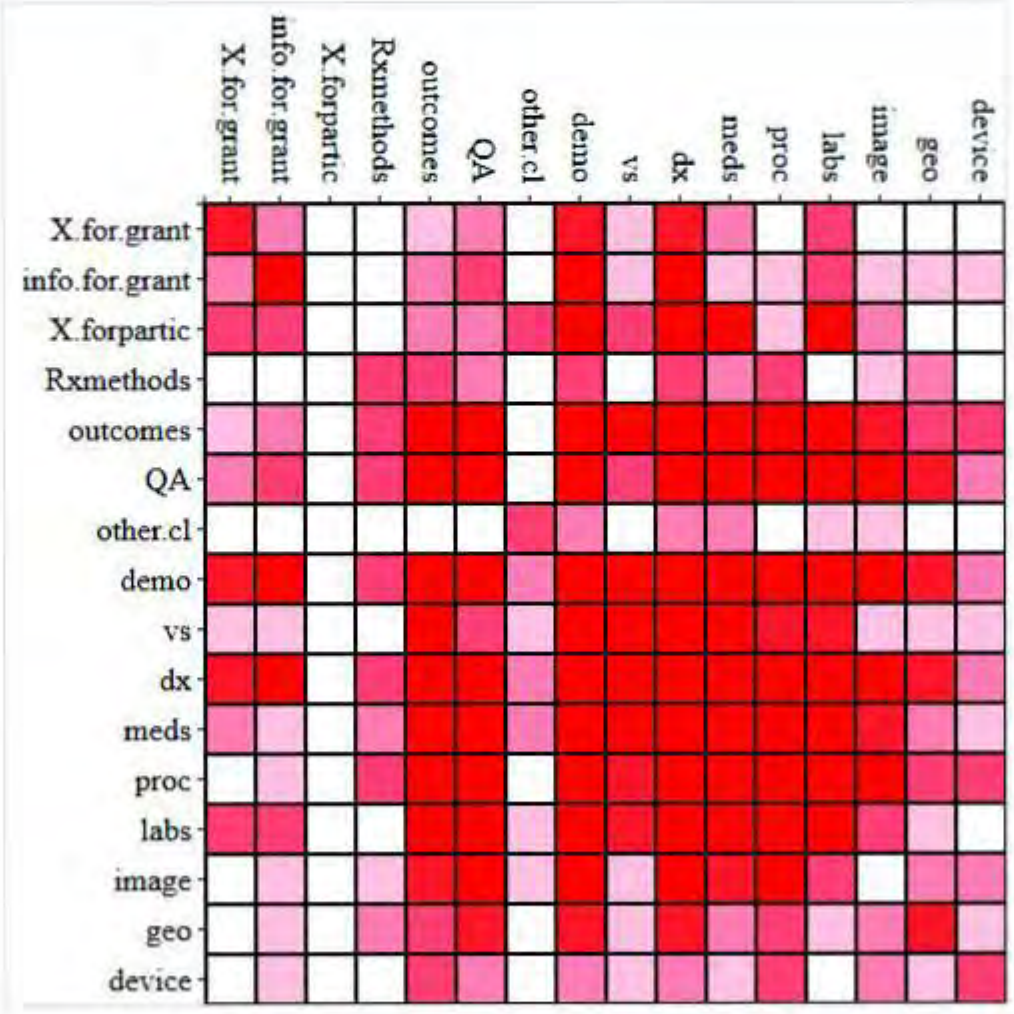
Parallel Sets Visualization



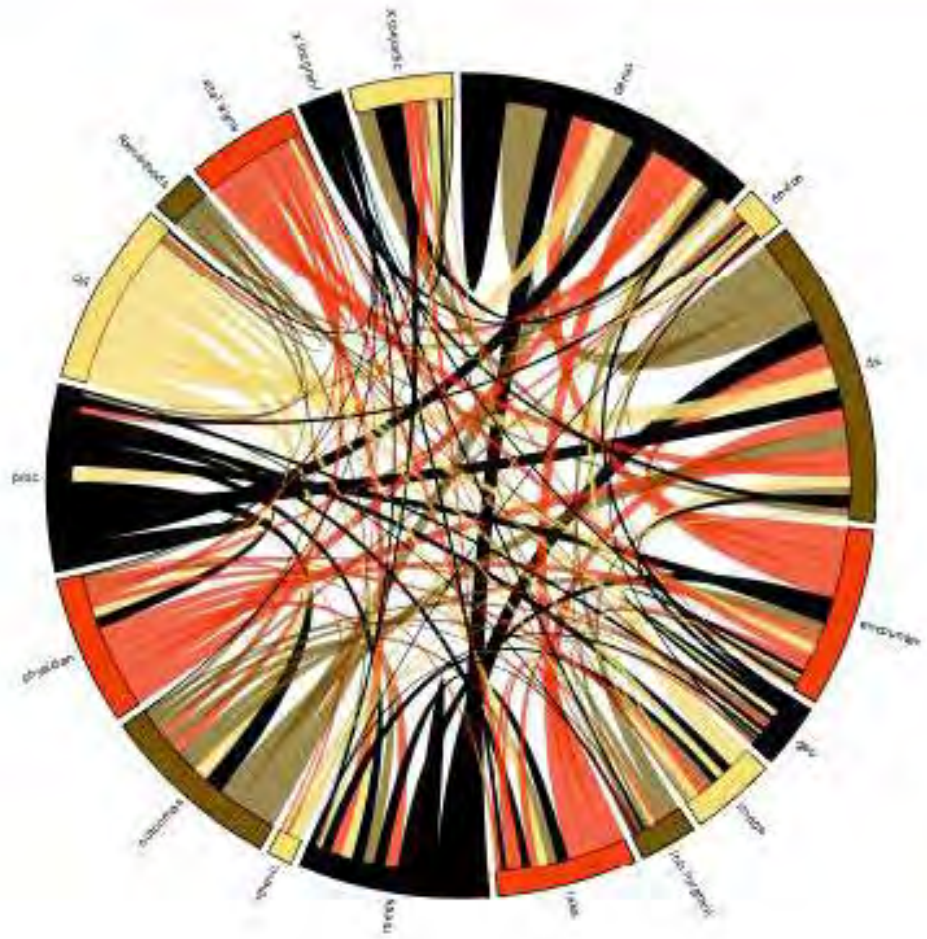
Parallel Sets Visualization

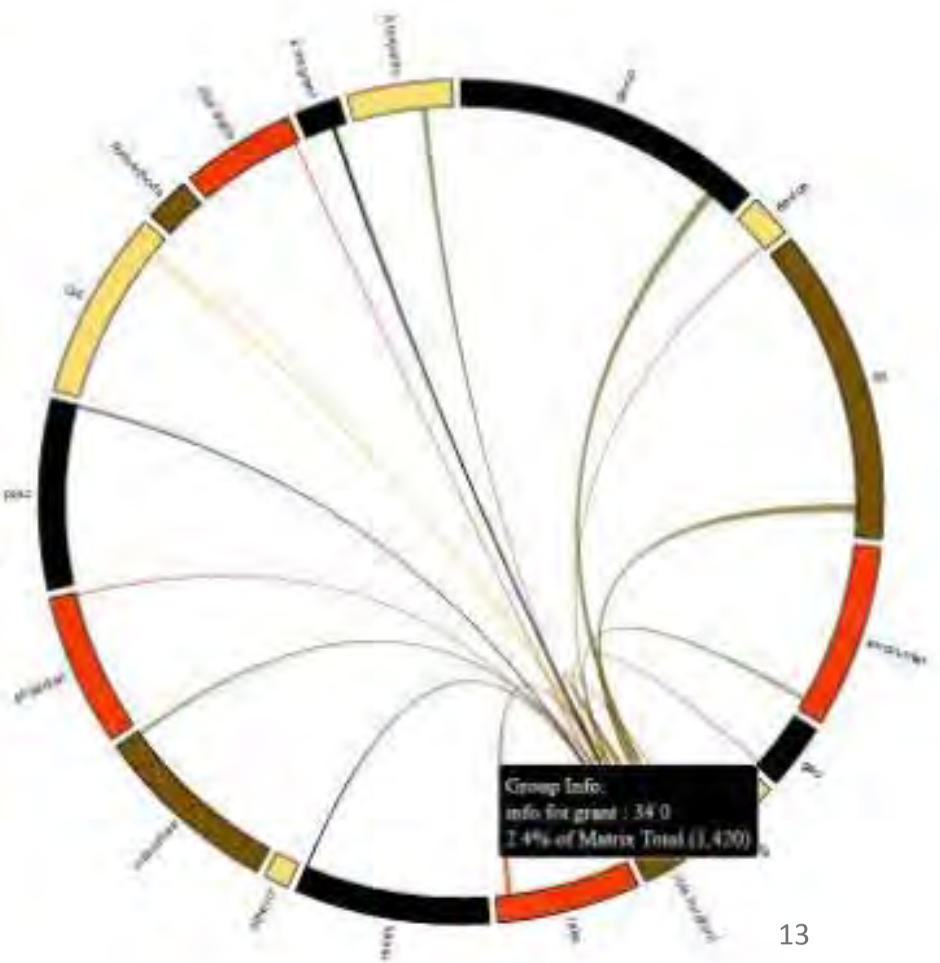


Co-occurrence Matrix



Chord Diagram





Future Work

- These visualizations are just some of the possible methods of multivariate visualization
- Explore further options
- Show visualizations to a variety of users

Acknowledgements

Dr. David Borland, RENCI

Dr. Vivian West, Duke University

Hina Shah, UNC Chapel Hill

Dr. Ed Hammond, Duke University

Acknowledgements

This work was supported by the US Army Medical Research and Material Command (USAMRMC) under Grant No. W81XWH-13-1-0061.

Thank You!